For Reference

NOT TO BE TAKEN FROM THIS ROOM

Ex libris
Universitatis
Albertaensis

QUAECUMQUE VERA

The University of Alberta
Printing Department
Edmonton, Alberta

Digitized by the Internet Archive
in 2023 with funding from
University of Alberta Library

https://archive.org/details/Wong1973

THE UNIVERSITY OF ALBERTA

AUTOMATIC TELEPHONE DIRECTORY ASSISTANCE

by

© ANDREW LAP-SANG WONG

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE

OF MASTER OF SCIENCE

DEPARTMENT OF COMPUTING SCIENCE

EDMONTON, ALBERTA

FALL, 1973

ABSTRACT

An investigation was made on the problem of automating the telephone directory assistance system. After a review on the computer methods of person identification, studies made directly toward automating the telephone directory assistance system were also introduced. A system based on a special coding method was studied and presented as a feasible solution to the problem. Also, possible extension to the system was discussed.

ACKNOWLEDGEMENTS

The author wishes to thank:

Dr. K. V. Leung, his thesis supervisor, for guidance throughout the course of study.

Professors I. N. Chen, T. R. Marsland and W. J. Myers for their many constructive comments on this study.

Professor B. M. Harden, for his continuous encouragement and final reading of the manuscript.

Department of Computing Science for the financial assistance and the Data Centre of the Northern Alberta Institute of Technology which facilitated the program testing portion of this study.

Mrs. L. Bowes and Mr. D. Brown, Department of Business Administration, Northern Alberta Institute of Technology, for reading the manuscript.

His wife, May, for her encouragement and help in typing the manuscript.

Miss C. Hergert, for her help in typing the manuscript.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF TABLES (continued)

<u>LIST OF DIAGRAMS</u>

Chapter I

GENERAL INTRODUCTION

Human handling of information has long been a problem area in information processing since the advent of modern digital computers. It is the slowest and most difficult to improve facet of a complete data processing system. Our point can best be illustrated by an example: In a cross-country airline reservation system, data are transferred rapidly between service centers and the processing center, search on file is done in the best-known fashion; yet most of the delays are caused either by the ticket agent or the customer. There are many factors that could contribute to the delays. For instance, slow keying (for CRT keyboard), communication problems between the customer and the ticket agent, indecision by the customer, or third party interruption of the ticket agent are among the most common reasons.

Obviously, if the human element can be eliminated from the system, information processing can be improved many-fold. Although complete absence of human operator seems impossible for most systems, the ultimate goal might be to achieve a minimal level of human intervention.

Telephone directory assistance is a classic case. A medium-sized telephone company serving a population of 500,000 has to employ hundreds of operators to answer calls. A bulky telephone book has to be updated every month for entering of new listings, deletion of old listings and changes of current listings. Expensive electronic equipment has to be purchased. Above all, enough space has to be provided for the equipment and operators. The process is slow and prone to error. The whole

operation is costly to set up and maintain.

Many attempts have been made to mechanize the operation of telephone directory assistance without much success. Heavy human involvement is retained because their function in the system is found to be "either not economical to mechanize or perhaps even impossible to fully mechanize".[2]

It is the purpose of this report to propose a telephone directory assistance system without human operators. By dialing the telephone on his desk, a caller makes his enquiry and receives a pre-recorded human voice answer. The technique involved is called "Computer-controlled message synthesis" and it was developed at Bell Labs in 1970.[1] The enquiry is recorded and coded into a search-record; this search-record is passed to a search program; search files are on-line; for example, disk-file. The result--a telephone number if the search is successful, otherwise a message is transferred to a simulator which performs the message synthesis.

Chapter II of this report gives a survey of person identification techniques, since telephone directory search is a typical person identification problem. In this survey six representative and significant studies are introduced.
The state-of-the-art in telephone directory assistance is discussed in Chapter III, while the local telephone company (Edmonton Telephones) is used as a study base. Also, current developments at Edmonton Telephones are reported for comparison.

The basis of this study is a special direct-dialed coding method on a telephone set (DD Code). Details and the justification of this method are examined in Chapter IV. The approach taken in this study

is clearly different from conventional approaches as reviewed in the previous two chapters. Also three of the name coding methods discussed in Chapter II (Blair's, Davidson's and SOUNDEX) were programmed with different sample sizes and the results are compared with DD Code on merits of discriminating power and degree of redundancy.

Chapter V describes the proposed telephone directory assistance system with detailed file organization and maintenance procedures. The problem of system performance measurement is also dealt with.

CHAPTER II

THE PROBLEM OF PERSON IDENTIFICATION

A. Pioneer Work

Some of the most important pioneer work in the field of identification of medical documents was done in Britain. In 1948, Lancelot Hogben, Muriel Johnstone and K. W. Cross[14] of University of Birmingham and Birmingham United Hospital were commissioned to look into the possibility of designing a system of medical documentation that could make provision for the efficient identification of individual patients. Their basic findings are listed below:

1. birth names are not specific (over a sixth of the entire population of England and Wales falls in one of the 50 most frequent surnames);
2. an initial proposal of a six-cipher code, using patient's birth date (day, month, year);
3. a ten-cipher code, the 4 additional ciphers (2 for surname, 2 for first name) being distributed to assure approximately equal relevance to each of the 10,000 compartments which 4 ten-row columns accommodate.

An initial analysis of the run of letters in 88,000 surnames of a Midland telephone directory was carried out and surnames were grouped into 100 blocks of equal frequency with a 2-figure code for each block. Two ciphers were used to code first names with the first digit distinguishing the sex of the individual.

Advantages of this method:

1. two individuals with same first names but different birth rank have a different code;
2. an initial of the first name will receive a different code than the full first name;
3. search on surname will have equal chance of a find in one of 100 blocks;
4. files can be arranged in birth date sequence;
5. unified patient numbering system;

A possible search scheme would be:

1. Arrange records in order of birth date, then name cipher; (birthdate ciphers should be arranged in increasing order of year, month and day.)
2. Search on birth date (if there are duplicates, then search on name ciphers).

B. <u>Phonetic Technique</u>

1. SOUNDEX[10] - The Phonetic Name Coding System

   SOUNDEX is a name coding system designed to solve some problems in name indexing. They are:

   a. different forms of spelling of the same or similar surname;
   b. errors in spelling;
   c. misinterpretation of handwriting;
   d. translation of names of foreign origin.

   The system is based on the principle that there are certain key letters (consonants) in the alphabets which cannot be eliminated from a proper name without making it

into something else. If we retain these letters in a name compression coding system, we have also retained the 'features' or 'characteristics' of the proper name.

Since names are usually filed as written and found as spoken, there is a problem in determining the exact apelling of a name. SOUNDEX* codes similar names or variations in spelling into one group.

The merits of the SOUNDEX system can best be demonstrated by the following examples:

Name	Code
MARAN	M650
MERAN	M650
MIRAN	M650
MORAN	M650
MOREN	M650
MORRAN	M650
MOURAN	M650

SOUNDEX has its shortcomings, too, particularly from the viewpoint of computer name coding methods. SOUNDEX does not utilize the advantages of speed and precision of a computer.

There are two major difficulties in record linkage by names:

a. names that should not be linked are linked;

b. names that should be linked are not linked.

*Note: see also Appendix C for rules of SOUNDEX.

The SOUNDEX method is especially weak in that names that are distinctly different are coded the same. The system is 'loose' in the sense that it cannot distinguish the finer features of names.

The following are some examples:

a. names to be distinguished by a vowel are coded the same in SOUNDEX. This is rather common with oriental names. For instance:

Name	Code
WONG	W520
WING	W520
WANG	W520

b. names with insufficient consonants will be coded with zeros which provides little discriminating power. For example:

Name	Code
HALL	H400
HILL	H400
HULL	H400
WU	W000
WA	W000
WEI	W000

c. most consonants are coded the same. For example: c, g, j, k, q, s, x, z all received the same code - code 2.

Name	Code
MORRIS	M620
MARK	M620
MERCOV	M620

d. names that should receive the same code are coded differently (some silent consonants). For example:

Name	Code
PSHEDEZKY	P322
SHEDEZKY	S322
SZCSYBALSKI	S222
SABALSKY	S142
MJELDE	M243
JELDEE	J430
JELDE	J430

In conclusion, the SOUNDEX system has its strong points and also its weaknesses. It is recognized that no system of name indexing is perfect in itself. It is also clear that the strong points of SOUNDEX are at the same time its weaknesses. For an ideal system, a combination of several methods may prove to be more satisfactory, because with the aid of a fast computer system, complex algorithms can be implemented and utilized speedily and efficiently.

2. Atomic Energy of Canada Ltd.

A group of researchers at the Biology Branch of the Atomic Energy of Canada Ltd. at Chalk River, Canada, also have done a fair amount of work in computer method of person identific-

ation. They were concerned mostly with vital record linkage.

In 1957, Newcombe, Axford and James[23] published a report describing the technique for co-ordinating from routine vital and health records information on heredity influences on health and for verifying the status for welfare program. Their pioneer effort was mainly on how to obtain reliable sources of information concerning the fine structure of family relationships from individual vital records.

To reduce the time to manipulate a considerable amount of name information, SOUNDEX coding method was used extensively. For example, full name code could consist of the following information (a total of 16 digits):

a. SOUNDEX code of father's surname;

b. SOUNDEX code of father's mother's maiden name;

c. SOUNDEX code of mother's maiden name;

d. SOUNDEX code of mother's mother's maiden name.

The advantages are obvious, either an operator can code-punch directly or a program can be set up to code-punch names.

Between 1957 and 1965, the Newcombe and Kennedy team of Atomic Energy of Canada published a number of articles (references 21-32) on computer methods of vital record linkage. To summarize their findings*:

a. they used the SOUNDEX coding method to facilitate the process of person identification;

*see also Appendix D on Factors Influence Choice of Identifiers

b. a weighted factor is used for different pieces of identifying information. Assuming the discriminating power of a particular item of identifying information depends upon its frequency of occurrence in the population, then a middle name initial of letter 'Z' has more discriminating power than the letter 'J'.

c. taking into consideration the combined discriminating power of all items of identifying information in the document, a greater degree of certainty can be achieved. In practice, they expressed the discriminating power for various agreements and disagreements of the different items of identifying information as logarithms so as to make them 'addable'. Tables of such values were prepared and listed in their report.

C. Letter Frequency Approach

Charles Blair of the Department of Defence, Washington, D.C. published his paper on 'A program for correcting spelling errors' in 1960 in Information and Control[6]. It is referred to as 'Blair's method'* in the following discussion.

By abbreviating names but retaining the 'kernel' of the original names, misspelled names which retain enough similarity to the original can be retrieved. Basically, Blair recognized that not all letters in a word are equally important. If the misspelled word happens to retain the important

*Note: Blair's method was programmed with various size samples for comparison purpose. See details in Chapter IV.

letters, it should receive the same abbreviation as the correctly spelled word.

The abbreviation algorithm is as follows:

1. score each letter of the name according to its frequency of occurrence in English text (a table of frequency was provided in his paper);
2. score each letter of the name according to its position (using a table which gives the logarithm of the desirability of deleting a letter as a function of its position).
3. total the scores for each letter;
4. take the four letters of lowest scores from left to right;
5. the four letters will be the 'abbreviation'.

It was found that names with small variations frequently resulted in the same abbreviation. One observation is that the choice of four as the size for abbreviation word might not be a wise one. In fact, the author of this thesis found 5 letters from the last name will give better discriminating power in comparison with just 4 letters. A similar problem was encountered in the project when attempting to determine the optimal word size allowed for the last name in order to minimize the duplicates of special codes and at the same time maximize the discriminating power*.

D. Scoring and Matching

There have been many attempts to deal with the problem of misspelled words during transmission of information. Miller and Friedman published in the Journal of Information and Control in 1958 on the subject of reconstruction of 'mutilated English Text'[7].

*Note: see Tables 4-13 and 4-14, discussion in Chapter IV, section G.

They claimed that the average person, given limited time to work, can correct passages reasonably well only if the mutilated text errors are less than 10%; the job is most difficult if it consists of random substitutions of wrong letters. Their approach is basically trial and error. According to the frequency of occurrence of each letter, substitutions are made. Their findings are not of direct concern to the name identification problem because:

1. English text was used as test data base. Word meaning in context can be taken into consideration, whereas person names do not have such built-in characteristics.
2. Their method is not readily programmable for a digital computer (no clearly defined algorithm).

E. Name Compression

Need arose for retrieval of misspelled names in Airlines Passenger Records. Davidson[13] (1962) tackled the problem by way of name compression. He avoided using the phonetic techniques for two reasons:

1. The international scope of names to be handled makes the phonetic equivalents of certain letters difficult to standardize.
2. The rapid turnover of airline agent personnel prohibits a system requiring training in phonetics.

He developed a spelling-matching technique which in one sense or another recognizes the 'essence' of a name despite the variant forms created by usual or unusual misspellings.

His compression scheme was in fact a well-known name coding

method published by IBM[15]. His search technique was also conventional: first a preliminary search on coded surname; if there is more than one match, then a search on coded full name; and eventually a display of matched records for manual check by an operator.

However, Davidson's research provided an interesting direction for further investigations. He derived a routine called "Ill-spelled Routine" for error recovery. Essentially, it is intended for spelling error correction and it is called for whenever spelling errors are more significant than 'vowel errors'. There are three rules used:

1. no letter appears in a code name more than twice;
2. space characters in coded surnames are packed at the right hand end;
3. repeated letters in a coded surname are not contiguous.

Also always look for a string of letters in the same sequence in both the coded surname and the retrieved record. This helps to suppress 'noises' caused by unmatched letters.

F. Universal Identifier (UID)

Person identification is a classical problem in information processing because the natural identifier (names) is a poor one in terms of uniqueness and discriminating power. There are just too many persons in the streets with the same names. Until a unique and universal identifier can be given to each individual, handling of information pertaining to human beings will continue to be a very difficult and frustrating job. Person name as an identifier although desirable, is ineffective and inefficient.

Attempts have been made many times in the last decade to establish a Universal Identifier (UID). Several European countries have already adopted some form of a UID system to facilitate processing of huge volumes of data about their citizens. This includes the Scandinavian countries and Great Britain. A number of others, like Japan and West Germanv are implementing similar systems.

The UID system of West Germany* is a twelve-digit number assigned to each citizen who is known officially to government by this twelve-digit number thereafter. To break down the 12 digit number:

1. six digits indicating birthday;
2. one digit for sex and the century of birth;
3. four digits to distinguish one from others born on the same day;
4. one digit for control purposes.

The Swedish UID** is composed of ten digits. The first six digits indicate the birthdate of the individual, then a three digit number to distinguish persons born on the same day (odd for men and even for women), plus a control digit. An earlier version of the Swedish UID system was introduced in 1947 and the control digit was added in 1968.

G. Conclusion

Many data processing specialists consider the universal use

* Note: TIME magazine, July 12, 1971.

**Note: See Appendix E for details.

of a UID will be the ultimate solution to many current problems in information processing. History shows disregard of the human aspect of a system is a common pitfall of most potentially great technological achievements. The fatal mistake was in designing a great system on paper without the adequate knowledge of the needs of prospective users.

Trying to fit human beings into a system rather than trying to fit a system to the needs of human beings has been costly for many well intended computer application systems. Failure to satisfy the needs of users is the fault of the system analyst, not the users. Some will argue that it is too time-consuming and costly to go all out and try to serve the users, but the shortcoming is of technology, not of human beings.

In view of the rapid advancement of computer technology, machines are built to work 10 times faster but relatively cheaper. The cost of data processing (speed and storage wise) has been reduced greatly over the last decade: e.g. The price of disk units was reduced by about one half.

The problem of person identification is not going to be solved by the introduction of UID, but rather by more sophisticated and human oriented system. UID is just a dream of 'lazy' data processing personnel. People just do not want to be identified by a number.

CHAPTER III

THE PROBLEM OF TELEPHONE DIRECTORY ASSISTANCE

A. Introduction

Telephone directory assistance has been provided by the telephone companies to their customers free of charge in the past. However, due to increasing demand and cost, telephone companies find systems used in the 60's can no longer cope with the current situation. Therefore, a much superior system is definitely needed. There are three areas which must be improved:

1. response time;
2. cost;
3. reliability.

In view of the above criteria, a computer-assisted directory system seems to be the most logical solution. Not only can it provide faster and more accurate answers to calls but printing of bulky directory 'updates' can also be eliminated.

There are at least four cities currently engaged in the study of computer-assisted directory systems:

1. Oakland, California;
2. New York, N. Y.;
3. Copenhagen, Denmark;
4. Edmonton, Alberta.

There is a prototype being tested in Oakland, California. The system features CRT terminals to display the 'hits'. 'Hits' are records that are found with the same major key--last name.

Last names are keyed in by the operator using the CRT keyboard. If there is more than one 'hit' displayed on the CRT screens, the terminal operator then asks the caller to supply further information (e.g., first name, address) that may help to identify the desired number. When the correct record is identified, the operator quotes the telephone number, (one number only). No numbers are to be given if the operator fails to find a single unique record.

B. Edmonton Telephones

Currently at Edmonton Telephones there are 48 stations in the Directory Assistance Department. There are 130 operators employed, working three full shifts, with an annual budget over $780,000.

The Directory Assistance Department is also responsible for giving such service to Northern Alberta as well. The most significant aspect of this service is that it's free. It was estimated that once a charging scheme was adopted, demand on such a service could drop to only 40% of the current rate. An extensive survey has been done by the Directory Assistance Department of Edmonton Telephones. There are several interesting findings:

1. currently, average response takes about 20 seconds;
2. it seems just four letters (three from the last name, one from the first initial) are sufficient to identify a person;
3. if there is more than one 'hit' during the search, more information is asked for (e.g., address);

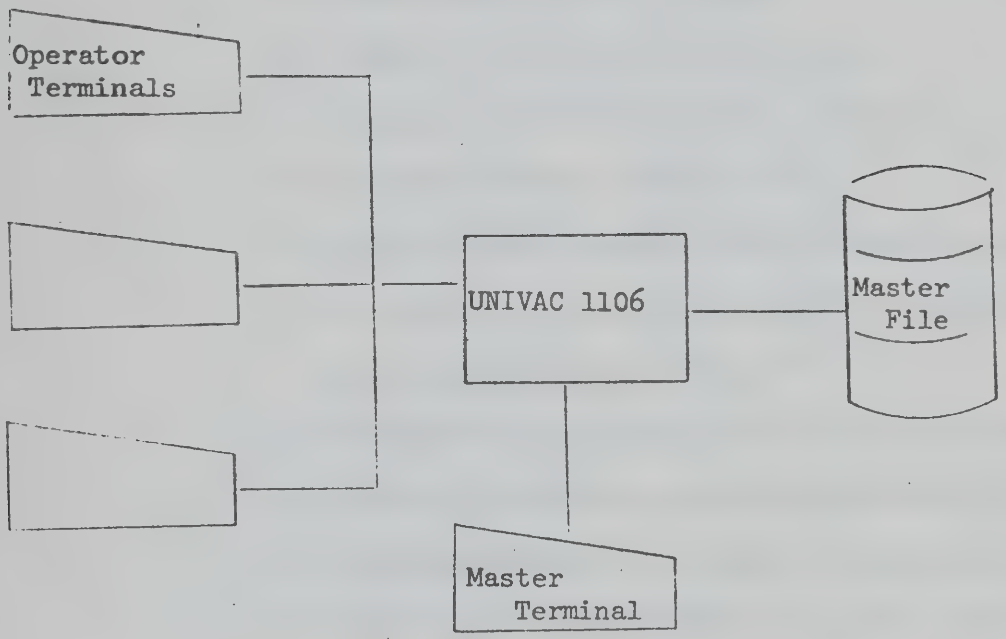
4. customer misspelling causes difficulties (as also do short names instead of full);
5. current equipment is not capable of handling the demand projected for 1974;
6. therefore Computer-assisted, directory search is deemed inevitable;
7. this service is used for business numbers 70% of the time;
8. directory assistance department policy is only one number to be given even when there are two possible numbers;
9. there is a 'most-frequently-called' number list (it is learned by word-of-month, and updated irregularly);
10. there are approximately 200,000 records on file at the present;

The Directory Assistance Department at Edmonton Telephones is actively engaged in the study of a computerized directory assistance system. The solution they are seeking would have the following features:

1. a long-term solution;
2. one that can cut response time by half(current manual system takes about ten seconds per call on the average);
3. possible reduction of total operating cost.

Edmonton Telephones is considering adopting Oakland's system with some modifications. They are still in the planning stage and expect completion in 1974.

Basically, the system visualized by Edmonton Telephones is as



HARDWARE CONFIGURATION

EDMONTON TELEPNONES

DIAGRAM 2 - 1

shown in Diagram 2-1.

The Master file is stored on a dedicated frum; the CRT terminals communicate to the UNIVAC 1106 main frame, and there is one master terminal for control purposes.

System characteristics:

1. human operators still needed;
2. 'hits' are displayed on CRT screen;
3. totally redesigned functional keyboards;
4. operator responds (human voice);
5. can also handle conventional intercept calls*;
6. two separate files are maintained - business numbers file, and residence numbers file;
7. to support approximately 40-50 CRT terminals;
8. printed statistics to be produced every 30 minutes for control purpose (e.g., number of calls, etc.);
9. three types of transactions - regular changes (correction) through a monitor terminal (master terminal), inter-terminal communication (to help handle difficult cases), update of the frequently-called numbers list;
10. response time at around ten seconds;
11. save on printing of telephone directory updates;
12. a maximum of 42 hits are accommodated, on 3 CRT screen pages (14 lines per screen page);
13. language used is FORTRAN V;
14. master file stored on a dedicated drum, on-line 24 hours.

* intercepts calls for numbers dialed which are not in service.

File shall be in alphabetical sequence with 3 fields in each record. Record format is basically the same as a printed telephone directory. The three fields are:

1. name: three letters from last name plus first name initial;
2. address: may adopt Assessor Department's Roll Number file (City of Edmonton) address coding method - the 4-3-2-1-5 method;

   home number: 4 3 2 1 5

   e.g. 10510 - Jasper Ave., Apt. 20

   coded as 10510/JASP/AVE/ /20

   12 Sir Winston Churchill Square

   coded as 12/SIR /WIN/CH/S
3. telephone number: seven digits number.

At the time this thesis is prepared Edmonton Telephone Directory Assistance System is still at its planning stage. Therefore, there is no further information available.

C. Bell Telephone System

Bell Telephones is perhaps the most active in the research into a better directory system to replace a manual system. Their past attempts like 'microcards' and 'microsticks' were not widely used because the records were expensive to update and the saving in time compared with printed directory was negligible.

In 1968, an Automatic Intercept Service invented by W. A. Winkleman[3] was announced by Bell Telephone Labs. The system is designed to answer calls to non-working numbers. Pre-recorded phrases and digits are put together to tell callers what numbers has been dialed, report status on the telephone and give a number

where the party can be reached. Similar systems were installed in 25 different cities.

R. D. Trupp[1], also of Bell Telephone Labs, published a paper in 1970 on computer-controlled message synthesis which confirms that a computer assembled message is feasible. Winkleman's system worked quite well with intercept calls because a search of a file for a phone number always produces a unique record. The significant result is output with simulated voice. Trupp's research seemed to be the answer.

The problem with any telephone directory assistance service is that a search with names (most readily available information from a caller) generally does not produce unique results. To request the phone number of Mr. John Smith would mean a possible choice out of too many. Therefore multiple keys are needed, and a multiple search has to be performed to uniquely identify a record. The same task performed by an operator takes time and is error-prone.

The research results of both Winkleman and Trupp are used as building blocks in the proposed Automatic Directory Assistance System in this report. The issue here is that if a computer can control and synthesize messages (no human operators), is it also possible for the computer to accept calls directly from the callers ? If this is possible, then a truly automatic system can be built. The theme of this study is to investigate the feasibility of directly dialed messages (names, address etc.) to be accepted and 'understood' by the computer. Details can be found in Chapters IV and V.

D. Rothrock's Proposal

H. I. Rothrock[2] published his doctoral thesis on Computer-assisted Directory Search in 1968. In his study, a model of man-machine interaction was constructed to illustrate a directory assistance system in which the optimal operator keying strategy was the prime concern. Since the employment of human operation is a major part of this system, all the disadvantages of such a system discussed above still apply. However, his detailed analysis of the telephone directory and the distribution of the descriptors (items of information to identify an individual listing), and the pattern of customer requests are of vital importance to this study. In fact, three significant points were of particular interest to our study:

1. in North America, any city of medium to large population would have similar distribution of family names;
2. on searching strategy, the use of any more than five letters from the last name does not increase the discriminating power significantly;
3. on customer request pattern
   a. over 70% are for business listings;
   b. of all descriptors from a directory, only four of them are likely to be given by a customer (listed name, next name or business type, house number, street number).

In Rothrock's study, file organization, maintenance and file update were not of prime concern. Rather, various keying strategies as well as patterns were examined and compared. Two optimal keying

strategies were devised:

1. for residential listings:

   a. 3LN + 3SN (preferred) which means the combination of the first 3 letters from the last name and the first 3 letters from the street name.

   b. 4LN + FI (if SN is not furnished) which means the combination of the first 4 letters of the last name and the first initial.

2. for business listings:

   a. 3FN + SN which means the first 3 letters from the finding name (first name of the company) plus the first letter of the street name.

Also, he proposed a search strategy, based upon the keying strategies he established. The concept of dividing power was defined and used as a measurement for descriptor-discriminating power.

The retrieval of listings is controlled by the computer in such a fashion that as the operator enters the codes, they are examined by the computer to determin the 'adequacy' for achieving a 'desirable' number of listings. This is referred to as 'AUTOSTART' in his thesis. The procedure assumed for his model is that the operator will continue to key descriptors according to general keying rules that either:

1. AUTOSTART occurs (15 or fewer listings will be returned and then displayed on the CRT);
2. no further descriptors available, START key is pressed.

Operator can strike the START key erroneously after an inad-

equate combination of descriptors has been keyed. To provide a kind of 'screening' function, 'AUTOSTQP' can be implemented for the purpose of minimizing time-consuming ineffectual searches. It also detects inadequate descriptor combinations and returns appropriate messages to the operator to indicate the problem. Implementation criteria for AUTOSTART and AUTOSTOP were discussed briefly in his report. For AUTOSTART:

1. establish a threshold for minimum number of listings;
2. file to be index sequential file to reduce search effort;
3. must accommodate changes and updates on files;
4. must be responsive to external conditions (e.g., traffic measurement can be taken into consideration during listings searches);
5. time spent on AUTOSTART decision must greatly reduce retrieval time;
6. must be economical to implement;
7. must not alienate human operators.

E. Conclusion

A survey of present studies in Directory Assistance Systems revealed a key common feature. That is, human operators are still actively engaged in functioning with the hardware and software system as an integral unit. The reasons were:

1. nobody ever considered a direct dialing method;
2. human intuition may help solve decision problems;
3. natural human voice output may be more acceptable to the public.

Whether these reasons are still valid today is debatable. In view of the fast pace of society and demand on the service of a Directory Assistance System, a breakthrough in design ought to anticipate a much improved system.

It is the purpose of this report to outline a Directory Assistance system which attempts to solve the problem without the intervention of a human operator and also to give improved efficiency and reliability. Input is accepted through dialing on a telephone and a human voice is simulated as output.

CHAPTER IV

DIRECT CODING OF ALPHANUMERIC INFORMATION VIA A TELEPHONE SET

The theme of this study is to design a telephone directory assistance system without human operators. There are basically three phases of the system:

A. All enquiries are dialed in on a standard telephone set (see Diagrams 4-1 and 4-2 of touch-tone and standard rotary dials)*, using the special characters * and # as message delimiters(note the proposed changes on the Diagrams to include the characters Q, Z and blank). In this way, names utilizing a 27-character alphabet are translated into codes utilizing 9 digits. The translation algorithm is in fact the standard telephone dial (plus the proposed changes), hence the name chosen for the system - Direct-Dialed Code or DD Code. The coded message is then passed to a search program.

B. The search program will do a generic search on an index ed sequential file (details in ChapterV). The result of the search - a phone number, or a message if no match - shall be passed on to a message-synthesizer[1].

C. A simulated human voice generated bv the message synthesizer will be heard by the caller over the telephone telling him

*Note: The proposed changes to the telephone dials shown in Diagram 3-1 and 3-2 (both touch-tone and rotary types), are to include the character Q,Z and blank which are absent on current dials. It is suggested key 1 should be used for these.

A touch button phone keyboard

(with proposed changes)

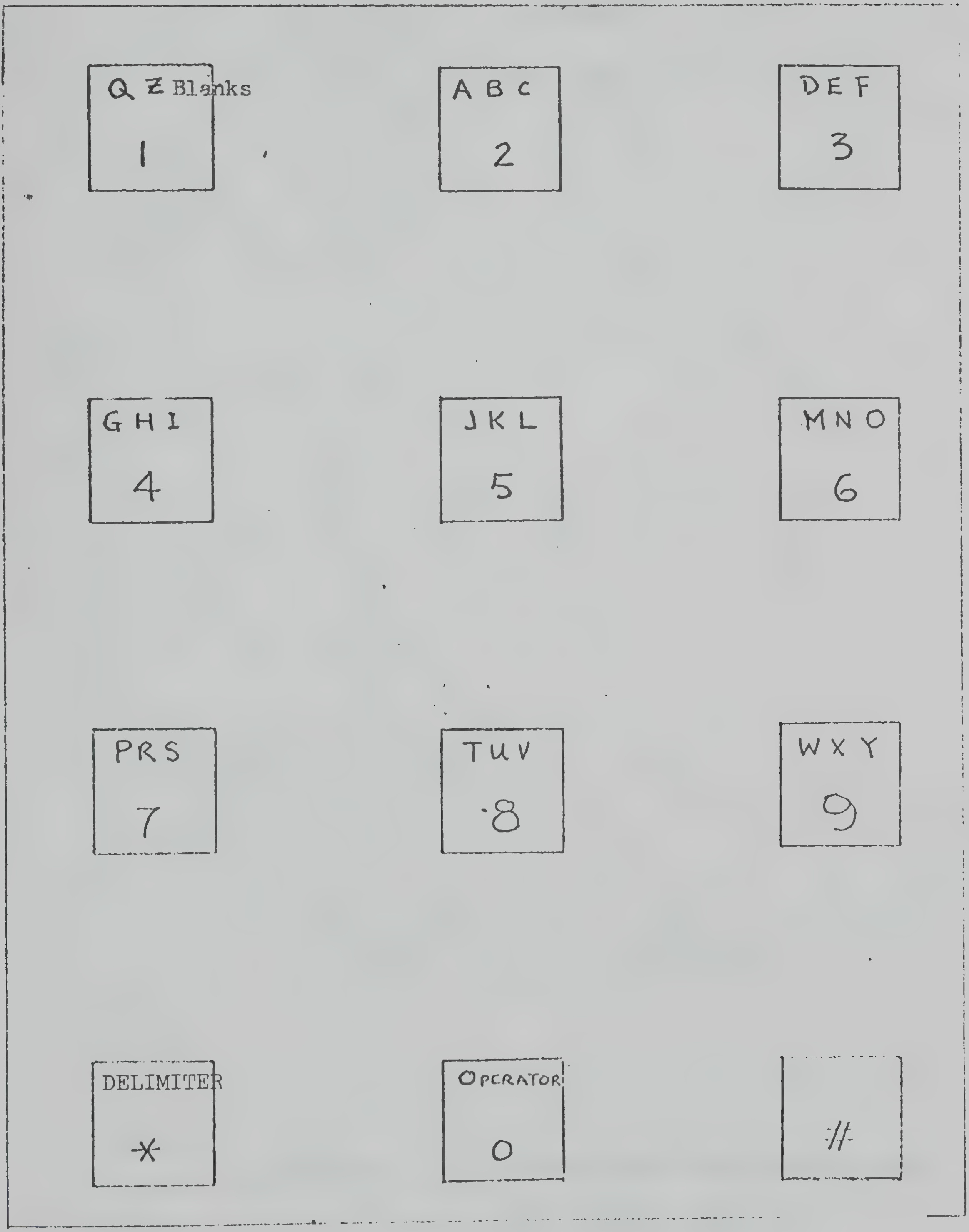


DIAGRAM 4 - 1

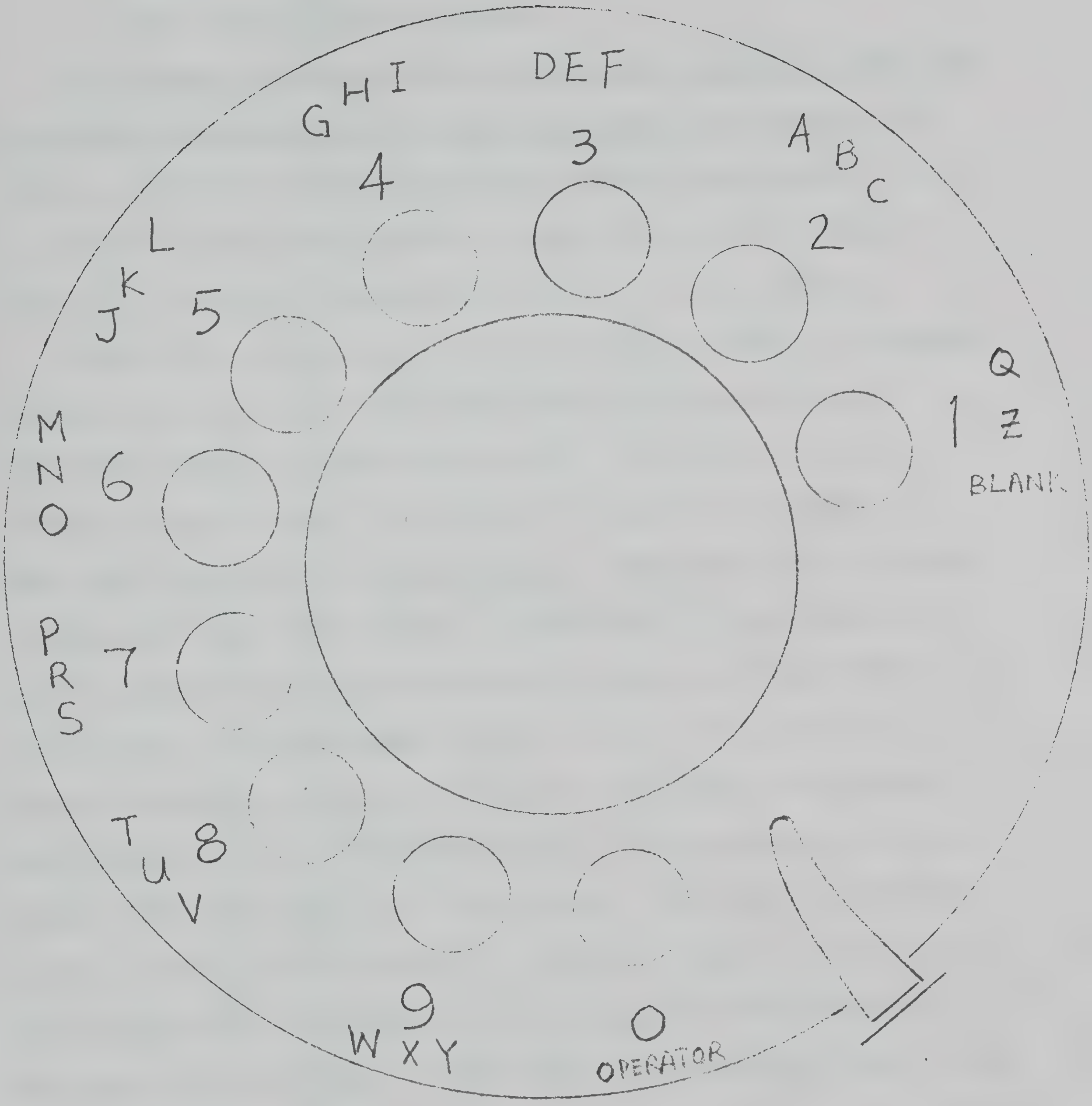


DIAGRAM 4-2 A Conventional Rotary Telephone Dial

(with proposed changes)

whether a phone number has been found with the given information, and the number if found.

This research is based on the fact that the alphabetic surname and its DD Code representation have a near one-to-one correspondence. No rigorous mathematical proof can be given; however, empirically, it can be demonstrated that the assumption is very close to truth (shown in Tables 4-1 to 4-12 and discussion in section C of this chapter).

Tables 4-1 to 4-12 are constructed to show the amount of redundancy introduced by using the DD Code rather than the alphabetic names. In each sample (size N), comparisons were made of different numbers of characters (first column - 'LAST NAME') to determine the number of codes that were duplicated (second column - 'SAME CODE') and the number of duplicated names (third column - 'SAME ALPH'). These, substracted from the sample size, gave respectively the number of UNIQUE CODES (sixth column) and UNIQUE NAMES (seventh column). The difference between the number of duplicated codes and the number of duplicated names (fourth column - 'DIFFERENCE SC-SA') is then expressed as a percentage of N, the sample size (fifth column - 'REDUNDANCY IN PERCENT') to give a measure of the increase in redundancy owing to the translation into the DD Code. The numbers of unique codes and unique names are also given as a percentage of sample size in the eighth and ninth columns respectively.

Thus, in Table 4-1, for a sample size of 2000, comparing on 10 characters (first row) gives an increase in redundancy of 0.95%, as also does a comparison on 9 characters. This redundancy increases rapidly to 4.25% for a comparison on 5 characters (bottom row).

TABLE 4-1 N = 2000 LAST NAME ONLY

LAST NAME	SAME CODE	SAME ALPH	DIFFERENCE SC -SA	REDUNDANCY IN PERCENT	UNIQUE CODE N -SC	UNIQUE NAME N - SA	UNIQUE CODES IN PERCENTAGE	UNIQUE NAMES IN PERCENTAGE
10	545	526	19	0.950	1455	1474	72.75	73.70
9	545	526	19	0.950	1455	1474	72.75	73.70
8	548	527	21	1.050	1452	1473	72.60	73.65
7	556	535	21	1.050	1444	1465	72.20	73.25
6	579	552	27	1.350	1421	1448	71.05	72.40
5	683	598	85	4.250	1317	1402	65.85	70.10

TABLE 4-2 N = 2000 LAST NAME AND FIRST NAME, 3 LETTERS

LAST NAME	SAME CODE	SAME ALPH	DIFFERENCE SC -SA	REDUNDANCY IN PERCENT	UNIQUE CODE N -SC	UNIQUE NAME N - SA	UNIQUE CODES IN PERCENTAGE	UNIQUE NAMES IN PERCENTAGE
10	224	222	2	0.100	1776	1778	88.80	88.90
9	224	222	2	0.100	1776	1778	88.80	88.90
8	224	222	2	0.100	1776	1778	88.80	88.90
7	226	224	2	0.100	1774	1776	88.70	88.80
6	232	227	5	0.250	1768	1773	88.40	88.65
5	260	236	24	1.200	1740	1764	87.00	88.20

TABLE 4-3   N = 6644   LAST NAME ONLY

LAST NAME	SAME CODE	SAME ALPH	DIFFERENCE SC -SA	REDUNDANCY IN PERCENT	UNIQUE CODE N -SC	UNIQUE NAME N - SA	UNIQUE CODES IN PERCENTAGE	UNIQUE NAMES IN PERCENTAGE
10	2071	1965	106	1.595	4573	4679	68.83	70.42
9	2073	1967	106	1.595	4571	4677	68.80	70.39
8	2083	1967	116	1.746	4561	4677	68.65	70.39
7	2130	2015	115	1.731	4514	4629	67.94	69.67
6	2292	2107	185	2.784	4352	4537	65.50	68.29
5	2884	2339	545	8.203	3760	4305	56.59	64.80

TABLE 4-4 N = 6644 LAST NAME AND FIRST NAME, 3 LETTERS

LAST NAME	SAME CODE	SAME ALPH	DIFFERENCE SC -SA	REDUNDANCY IN PERCENT	UNIQUE CODE N -SC	UNIQUE NAME N - SA	UNIQUE CODES IN PERCENTAGE	UNIQUE NAMES IN PERCENTAGE
10	112	112	0	0.000	6532	6532	98.31	98.31
9	112	112	0	0.000	6532	6532	98.31	98.31
8	113	112	1	0.015	6531	6532	98.30	98.31
7	113	113	0	0.000	6531	6531	98.30	98.30
6	114	114	0	0.000	6530	6530	98.28	98.28
5	117	114	3	0.045	6527	6530	98.24	98.28

TABLE 4-5 N = 6619 LAST NAME ONLY

LAST NAME	SAME CODE	SAME ALPH	DIFFERENCE SC -SA	REDUNDANCY IN PERCENT	UNIQUE CODE N -SC	UNIQUE NAME N - SA	UNIQUE CODES IN PERCENTAGE	UNIQUE NAMES IN PERCENTAGE
10	2025	1920	105	1.586	4594	4699	69.41	70.99
9	2027	1922	105	1.586	4592	4697	69.38	70.96
8	2037	1922	115	1.737	4582	4697	69.22	70.96
7	2084	1970	114	1.722	4535	4649	68.51	70.24
6	2245	2062	183	2.765	4374	4557	66.08	68.85
5	2836	2292	533	8.128	3789	4327	57.24	65.37

TABLE 4-6   N = 6619   LAST NAME AND FIRST NAME, 3 LETTERS

LAST NAME	SAME CODE	SAME ALPH	DIFFERENCE SC -SA	REDUNDANCY IN PERCENT	UNIQUE CODE N -SC	UNIQUE NAME N - SA	UNIQUE CODES IN PERCENTAGE	UNIQUE NAMES IN PERCENTAGE
10	114	112	2	0.030	6505	6507	98.28	98.31
9	114	112	2	0.030	6505	6507	98.28	98.31
8	115	112	3	0.045	6504	6507	98.26	98.31
7	115	113	2	0.030	6504	6506	98.26	98.29
6	116	114	2	0.030	6503	6505	98.25	98.28
5	120	114	6	0.091	6499	6505	98.19	98.28

TABLE 4-7 N = 5856 LAST NAME ONLY

LAST NAME	SAME CODE	SAME ALPH	DIFFERENCE SC -SA	REDUNDANCY IN PERCENT	UNIQUE CODE N -SC	UNIQUE NAME N - SA	UNIQUE CODES IN PERCENTAGE	UNIQUE NAMES IN PERCENTAGE
10	1844	1756	88	1.503	4012	4100	68.51	70.01
9	1849	1761	88	1.503	4007	4095	68.43	69.93
8	1856	1761	95	1.622	4000	4095	68.31	69.93
7	1891	1791	100	1.708	3965	4065	67.71	69.42
6	2031	1879	152	2.596	3825	3977	65.32	67.91
5	2605	2108	497	8.487	3251	3748	55.52	64.00

TABLE 4-8 N = 5856 LAST NAME AND FIRST NAME, 3 LETTERS

LAST NAME	SAME CODE	SAME ALPH	DIFFERENCE SC -SA	REDUNDANCY IN PERCENT	UNIQUE CODE N -SC	UNIQUE NAME N - SA	UNIQUE CODES IN PERCENTAGE	UNIQUE NAMES IN PERCENTAGE
10	99	98	1	0.017	5757	5758	98.31	98.33
9	99	98	1	0.017	5757	5758	98.31	98.33
8	100	99	1	0.017	5756	5757	98.29	98.31
7	100	99	1	0.017	5756	5757	98.29	98.31
6	101	99	2	0.034	5755	5757	98.28	98.31
5	107	102	5	0.085	5749	5754	98.17	98.26
4	121	110	11	0.188	5735	5746	97.93	98.12

TABLE 4-9   N =10038   LAST NAME ONLY

LAST NAME	SAME CODE	SAME ALPH	DIFFERENCE SC -SA	REDUNDANCY IN PERCENT	UNIQUE CODE N -SC	UNIQUE NAME N - SA	UNIQUE CODES IN PERCENTAGE	UNIQUE NAMES IN PERCENTAGE
10	3947	3754	193	1.923	6091	6284	60.68	62.60
9	3951	3757	194	1.933	6087	6281	60.64	62.57
8	3969	3774	195	1.943	6069	6264	60.46	62.40
7	4040	3833	207	2.062	5998	6205	59.75	61.82
6	4357	3999	358	3.566	5681	6039	56.59	60.16
5	5417	4399	1018	10.141	4621	5639	46.04	56.18
4	7716	5177	2539	25.294	2322	4861	23.13	48.43

TABLE 4-16 N =10038 LAST NAME AND FIRST NAME, 3 LETTERS

LAST NAME	SAME CODE	SAME ALPH	DIFFERENCE SC -SA	REDUNDANCY IN PERCENT	UNIQUE CODE N -SC	UNIQUE NAME N - SA	UNIQUE CODES IN PERCENTAGE	UNIQUE NAMES IN PERCENTAGE
10	290	285	5	0.050	9748	9753	97.11	97.16
9	290	285	5	0.050	9748	9753	97.11	97.16
8	291	286	5	0.050	9747	9752	97.10	97.15
7	291	287	4	0.040	9747	9751	97.10	97.14
6	296	296	0	0.000	9742	9742	97.05	97.05
5	308	297	11	0.110	9730	9741	96.93	97.04
4	343	308	35	0.349	9695	9730	96.58	96.93

TABLE 4-11 N =14556 LAST NAME ONLY

LAST NAME	SAME CODE	SAME ALPH	DIFFERENCE SC -SA	REDUNDANCY IN PERCENT	UNIQUE CODE N -SC	UNIQUE NAME N - SA	UNIQUE CODES IN PERCENTAGE	UNIQUE NAMES IN PERCENTAGE
10	6662	6355	307	2.109	7894	8201	54.23	56.34
9	6674	6366	308	2.116	7882	8190	54.15	56.27
8	6698	6389	309	2.123	7858	8167	53.98	56.11
7	6818	6474	344	2.363	7738	8082	53.16	55.52
6	7292	6723	569	3.909	7264	7833	49.90	53.81
5	8879	7314	1565	10.752	5677	7242	39.00	49.75
4	11946	8351	3595	24.698	2610	6205	17.93	42.63

TAPLE 4-12 N =14556 LAST NAME AND FIRST NAME, 3 LETTERS

LAST NAME	SAME CODE	SAME ALPH	DIFFERENCE SC -SA	REDUNDANCY IN PERCENT	UNIQUE CODE N -SC	UNIQUE NAME N - SA	UNIQUE CODES IN PERCENTAGE	UNIQUE NAMES IN PERCENTAGE
10	707	700	7	0.048	13849	13856	95.14	95.19
9	707	700	7	0.048	13849	13856	95.14	95.19
8	708	701	7	0.048	13848	13855	95.14	95.18
7	708	702	6	0.041	13848	13854	95.14	95.18
6	713	705	8	0.055	13843	13851	95.10	95.16
5	729	713	16	0.110	13827	13843	94.99	95.10
4	769	729	40	0.275	13787	13827	94.72	94.99

A. DD Code coding method - (Direct-Dialed Code)

For any coding method, there are two criteria to be considered:

1. same code generated for same names.

2. different codes generated for different names.

Mathematically, it is referred to as one-to-one correspondence. The coding method proposed here --- Direct Dialed Code, utilizing the standard telephone dial with minor changes --- will generate the same code for the same name, but sometimes, different names may receive the same code (see Appendix F for examples). In fact, this is a common characteristic with all existing name compression and name coding methods. However, finding a coding method which has a one-to-one mapping relation between the code and the datum might not be all that important if it can be controlled. That is, the focal point of investigation should be on how to control these undesirable'duplications' or'redundancy' to a level that is feasible and efficient to work with, hence allowing practical applications of the coding method.

After a review of some of the representative name coding and name-compression methods currently being employed or seriously studied, an important fact comes to light - that no code compression technique as yet can generate unique codes. DD Code has the same disadvantage, but since these 'duplications' only occurred on a small percentage basis, as shown in Tables 4-1 to 4-12; it is still feasible and also very practical to use such a coding scheme for person identification purposes.

B. Advantages of a System Using DD Codes

Files could be organized into two main files (an R file for

residential and an B file for business listings), which would further reduce search time and errors. For a business file, the most powerful single item of information is the business type, because hardly any company will use another's name in the same line of business. However, since there is, at present, no workable business classification scheme available, only the name and address details are considered here.

The advantages of such a system are obvious:

1. minimal human intervention (no operators);
2. fewer errors (minimal manual handling of information);
3. no 'hold-ups' on the system while the caller is thinking or keying information;
4. reduces social problems (e.g. undesirable calls directed to the operators);
5. faster response (search time greatly reduced);
6. easy update on files (computer files);
7. no handling of bulky telephone books;
8. eliminate possible human conflicts (impatient callers, or operator in a bad mood);
9. space saving (compact machinery vs. operator stations currently used);
10. automatic charge accounting(charging for this service is inevitable in the future);
11. possible implementation of a cross-country directory assistance system network.
12. easy implementation, as a telephone set is available in most North American homes.

C. Uniqueness of DD Codes

To justify this claim, studies were carried out on several different files to examine the uniqueness of DD Code generation (or on the other hand examine the degree of duplications). Ideally, in order to test the validity and efficient application of DD Code technique, a telephone directory of a medium to large population should be used as data base. However, since there is none readily available for machine input and the task of creating one is too costly and time-consuming, random samples of name files were used instead. The findings in each case are discussed below:

1. Edmonton Telephone Directory

A survey on the names listed in the Edmonton telephone directory revealed a very interesting fact. The most common family name (Smith) has nearly 1000 listings (933) both residential and business. With DD Code, no other family names can generate the same code as Smith. A test program was written for this purpose. The 10 most popular English family names were tested and the result is convincing. Only one of them (Brown) can generate a code that has 'duplicates'. That is, the name Crown and the name Brown shall be coded the same using DD coding method. However, in view of the infrequency of occurrence of the name Crown (there is only one residential listing of Crown) this 'duplication' has very little effect on the efficiency of search on files coded by DD method. That means even though the one-to-one correspondence between names and their DD Codes breaks down in certain cases, the

effect on search efficiency of the system is negligible. Furthermore, if the caller can furnish more information other than just the family name (first name, address) to increase the discriminating power, this problem can be overcome easily.

2. Patient File

A file of 2,000 patient names was used as a data base to test the discriminating power of DD coding method. Since the names were extracted from a local hopsital, and the only reason these names were on this file was because they were sick once and were treated at that hospital, it is random enough as far as name variations go. Therefore, we can assume it represents a good cross-section of the different family names of the city and a good random sample from the telephone directory. The result shows with 2,000 names data base there are only 19 'duplicates', if only last names were used, which is about 0.95% from the total, as shown in Table-1. Even if we allow seven letters of the last name only, there are 21 duplicates and this represents a 1.05% of 2,000 names.

Checking for 'duplicates' using both last name and first name in DD code is undesirable because first name is not available for most of the listings.

3. Student Application File

In the study on student application file, there were a total of 6644 applications. But because some students filed more than one application to different departments some 'housekeeping' has to

be done on the file. That is, duplicate applications are to be deleted from the file. Birthdays of applicants with same last names and first names are checked. If birthdays are the same, it is highly unlikely they are two different persons. New Count is about 6440. That is a significant step in this study because the inclusion of these duplicate applications will certainly increase the count on 'duplicates', thus presenting a false picture of the discriminating power of DD Codes.

Again, we can use similar arguments. Since student applications consist of a good cross-section of names of the city (even for the province), we can assume it is a good random sample from the telephone directory. Using only last name DD codes, there are 106 'duplicates' or about 1.6% of probability of getting duplicates. The surprising fact is that, using both last name and first name DD Code, there are no 'duplicates'. Even if we reduce the number of letters of the last name DD code to seven letters code there are only 115 'duplicates' or a probability of 1.73%. The most significant result was found when using first initial DD code with last name DD code. There are no 'duplicates' even if using only 6 letters of last name and first initial in DD codes(Tables 4-3, and 4-4).

To summarize the above discussions, the problem of 'duplicates' caused by the use of DD coding method is really insignificant. Effectively, what will happen is that under DD coding a telephone directory will consist of fewer family groups but more 'family members'. In the case of the student application file, with alphabetic names, 71.4% of the total is of different family names; wh-

ereas after DD coding only 68.82% of the total is of different names. This shows a shift of only 2.65%.

Another interesting fact is that by comparing the patient file and the student file, it is found that with about 3.3 times increase of file size, the increase of duplicates is about 1.6 times. This suggested that although 2000 is not sufficient to cover most family name variations, the increase of 'duplicates' is not in linear proportion to sample size. Different student application files of 6619 and 5856 entries were also tested and the results was compared with those collected from 6644 entries; the results are very close, which suggests that a sample size over 5000 would give stable and unbiased results. This stability aspect of the redundancy shall be discussed in more detail in section F of this Chapter.

4. Student Registration File

In order to obtain a large sample for listing of the redundancy of DD Code, student registration files for four years were merged together to form a single student registration file. This size is 10038 records. Data retained are last name, first name, and corresponding DD code. Further, this file was merged with all the student application mentioned in Section C.3 above to form an even larger file with 14556 records. Steps were taken to ensure that no duplicate records appeared in this file, thus preventing inaccurate redundancy measurements.

The Edmonton telephone directory has approximately 200,000 listings, of which fewer than 2/3 are residential, giving about 140,000 residential listings in a city of half a million. We feel

a sample size of 14556 is large enough for testing purposes.

Redundancy introduced by the DD Code based on these two samples was studied in the same manner as above, as shown in Table 4-9 to 4-12. Redundancy for last name only is 2.1% with a sample size of 14556 and is 0.048% for last name and first three letters of the first name.

D. Optimal Size of Last Name

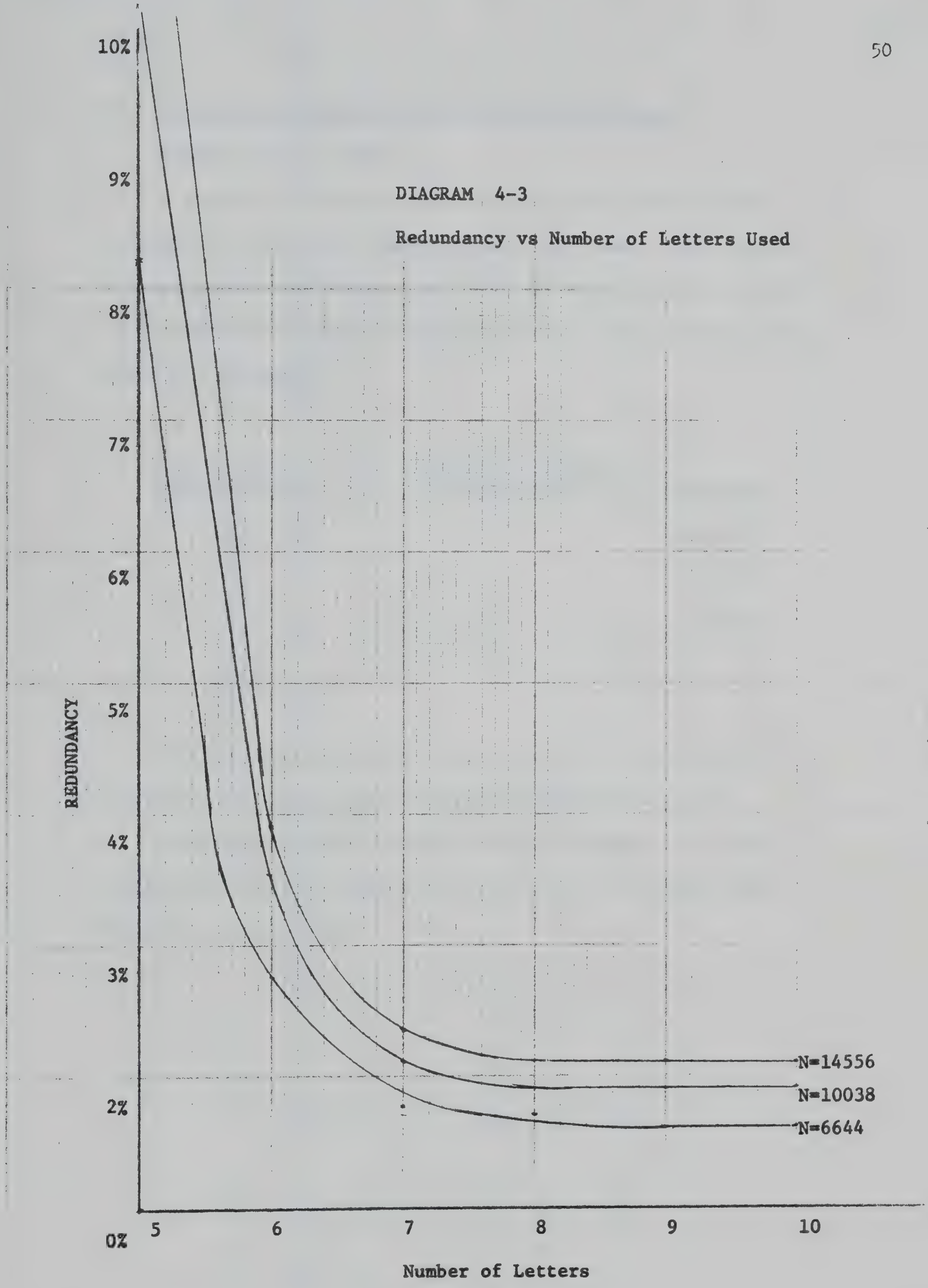
To study the optimal size of name for use with the DD Code, a diagram was prepared (Diagram 4-3) to show the relation between redundancy and the number of letters used. The three curves represent three different sample sizes, namely 14556, 10038 and 6644.

It is obvious that by increasing the size of last name in DD Code we are minimizing the number of 'duplicates'. However, it is evident from Diagram 4-3, that once more than seven letters are allowed for last name, an additional letter does not reduce the number of 'duplicates' significantly. In fact, seven letters gives approximately the same discriminating power as ten letters. Therefore, the optimal size for last name DD coding may be set at 7 coded digits*.

*Note: Most name coding study showed 6 letters to be the optimal size. Reference 16, Identification Techniques, IBM publication.

DIAGRAM 4-3

Redundancy vs Number of Letters Used



E. Choice of Letters from Last Name and First Name

(Tables 4-15 to 4-20)

A study of the selection of letters showed that the best strategy is to use both last name and first name. For example, data extracted from Tables 4-18, 4-19 and 4-20, using a total of 11 letters with different combinations from first and last names gives the following:

Number of letters from Last Name	Number of letters from First Name	Redundancy
10	1	0.309%
9	2	0.103%
8	3	0.048%

In fact, using only six letters from last name with three letters from first name only increases redundancy to 0.055%.

Therefore, the best letter selection strategy is to use 8 letters from the last name and 3 letters from the first name using DD coding method.

TABLE 4-13 N =10038 Blair's Method, Selected Letters from Last Name

LAST NAME	SAME CODE	SAME ALPH	DIFFERENCE SC -SA	REDUNDANCY IN PERCENT	UNIQUE CODE N -SC	UNIQUE NAME N - SA	UNIQUE CODES IN PERCENTAGE	UNIQUE NAMES IN PERCENTAGE
10	3769	3754	15	0.149	6269	6284	62.45	62.60
9	3770	3757	13	0.130	6268	6281	62.44	62.57
8	3775	3774	1	0.010	6263	6264	62.39	62.40
7	3803	3833	-30	-0.299	6235	6205	62.11	61.82
6	3882	3999	-117	-1.166	6156	6039	61.33	60.16
5	4190	4399	-209	-2.082	5848	5639	58.26	56.18
4	5173	5177	-4	-0.040	4865	4861	48.47	48.43

TABLE 4-14 N =14556 Blair's Method, Selected Letters from Last Name

LAST NAME	SAME CODE	SAME ALPH	DIFFERENCE SC -SA	REDUNDANCY IN PERCENT	UNIQUE CODE N -SC	UNIQUE NAME N - SA	UNIQUE CODES IN PERCENTAGE	UNIQUE NAMES IN PERCENTAGE
10	6376	6355	21	0.144	8180	8201	56.20	56.34
9	6379	6366	13	0.089	8177	8190	56.18	56.27
8	6390	6389	1	0.007	8166	8167	56.10	56.11
7	6435	6474	-39	-0.268	8121	8082	55.79	55.52
6	6556	6723	-167	-1.147	8000	7833	54.96	53.81
5	7030	7314	-284	-1.951	7526	7242	51.70	49.75
4	8488	8351	137	0.941	6068	6205	41.69	42.63

TABLE 4-15 N =10038 DD Code, Last Name with First Initial

LAST NAME	SAME CODE	SAME ALPH	DIFFERENCE SC -SA	REDUNDANCY IN PERCENT	UNIQUE CODE N -SC	UNIQUE NAME N - SA	UNIQUE CODES IN PERCENTAGE	UNIQUE NAMES IN PERCENTAGE
10	1520	1494	26	0.259	8518	8544	84.86	85.12
9	1520	1494	26	0.259	8518	8544	84.86	85.12
8	1524	1497	27	0.269	8514	8541	84.82	85.09
7	1531	1504	27	0.269	8507	8534	84.75	85.02
6	1572	1527	45	0.448	8466	8511	84.34	84.79
5	1718	1587	131	1.305	8320	8451	82.89	84.19
4	2054	1696	358	3.566	7984	8342	79.54	83.10

TABLE 4-16 N =10038 DD Code, Last Name with First Two Letters of First Name

LAST NAME	SAME CODE	SAME ALPH	DIFFERENCE SC -SA	REDUNDANCY IN PERCENT	UNIQUE CODE N -SC	UNIQUE NAME N - SA	UNIQUE CODES IN PERCENTAGE	UNIQUE NAMES IN PERCENTAGE
10	563	554	9	0.090	9475	9484	94.39	94.48
9	563	554	9	0.090	9475	9484	94.39	94.48
8	566	556	10	0.100	9472	9482	94.36	94.46
7	568	559	9	0.090	9470	9479	94.34	94.43
6	578	565	13	0.130	9460	9473	94.24	94.37
5	613	584	29	0.289	9425	9454	93.89	94.18
4	688	607	81	0.807	9350	9431	93.15	93.95

TABLE 4-17 N =10038 DD Code, Last Name with First Three Letters of First Name

LAST NAME	SAME CODF	SAME ALPH	DIFFERENCE SC -SA	REDUNDANCY IN PERCENT	UNIQUE CODE N -SC	UNIQUE NAME N - SA	UNIQUE CODES IN PERCENTAGE	UNIQUE NAMES IN PERCENTAGE
10	290	285	5	0.050	9748	9753	97.11	97.16
9	290	285	5	0.050	9748	9753	97.11	97.16
8	291	286	5	0.050	9747	9752	97.10	97.15
7	291	287	4	0.040	9747	9751	97.10	97.14
6	296	290	6	0.060	9742	9748	97.05	97.11
5	308	297	11	0.110	9730	9741	96.93	97.04
4	343	308	35	0.349	9695	9730	96.58	96.93

TAPLE 4-18   N =14556   DD Code, Last Name with First Initial

LAST NAME	SAME CODE	SAME ALPH	DIFFERENCE SC -SA	REDUNDANCY IN PERCENT	UNIQUE CODE N -SC	UNIQUE NAME N - SA	UNIQUE CODES IN PERCENTAGE	UNIQUE NAMES IN PERCENTAGE
10	2815	2770	45	0.309	11741	11786	80.66	80.97
9	2817	2772	45	0.309	11739	11784	80.65	80.96
8	2823	2777	46	0.316	11733	11779	80.61	80.92
7	2834	2766	68	0.467	11722	11790	80.53	81.00
6	2898	2823	75	0.515	11658	11733	80.09	80.61
5	3120	2911	209	1.436	11436	11645	78.57	80.00
4	3560	3054	506	3.476	10996	11502	75.54	79.02

TABLE 4-19 N =14556 DD Code, Last Name with First Two Letters of First Name

LAST NAME	SAME CODE	SAME ALPH	DIFFERENCE SC -SA	REDUNDANCY IN PERCENT	UNIQUE CODE N -SC	UNIQUE NAME N - SA	UNIQUE CODES IN PERCENTAGE	UNIQUE NAMES IN PERCENTAGE
10	1196	1181	15	0.103	13360	13375	91.78	91.89
9	1197	1182	15	0.103	13359	13374	91.78	91.88
8	1200	1184	16	0.110	13356	13372	91.76	91.87
7	1201	1186	15	0.103	13355	13370	91.75	91.85
6	1214	1193	21	0.144	13342	13363	91.66	91.80
5	1256	1216	40	0.275	13300	13340	91.37	91.65
4	1363	1250	113	0.776	13193	13306	90.64	91.41

TABLF 4-20 N =14556 DD Code, Last Name with First Three Letters of First Name

LAST NAME	SAME CODE	SAME ALPH	DIFFERENCE SC -SA	REDUNDANCY IN PERCENT	UNIQUE CODE N -SC	UNIQUE NAME N - SA	UNIQUE CODES IN PERCENTAGE	UNIQUE NAMES IN PERCENTAGE
10	707	700	7	0.048	13849	13856	95.14	95.19
9	707	700	7	0.048	13849	13856	95.14	95.19
8	708	701	7	0.048	13848	13855	95.14	95.18
7	708	702	6	0.041	13848	13854	95.14	95.18
6	713	705	8	0.055	13843	13851	95.10	95.16
5	729	713	16	0.110	13827	13843	94.99	95.10
4	769	729	40	0.275	13787	13827	94.72	94.99

F. Sample Size and its Effect on DD Code Redundancy

As expected, redundancy of DD Code varies in positive proportion to sample size. As shown in Table 4-21 (compiled from* Tables 4-1 to 4-20, in summarized form), using 10 letters; redundancy of DD Code incurred from 0.95% to 2.11% (about double), whereas sample size increased from 2000 to 14556 (7 times more). An attempt to calculate the optimal redundancy rate was abandoned due to insufficient data. However, by examining the rate of increase of the redundancy measure the maximum redundancy was estimated to be approximately 2.5%. We can expect the redundancy rate to become stable because variation on family name become stable. According to H.I. Rothrock[2] , any city of medium to large population in North America would have similar distribution of family names.

*Note: Diagram 4-4 show same data as Table 4-21, in graphical form.

SAMPLE SIZE / LETTERS	2000	5856	6619	6644	10038	14556
10	0.95	1.50	1.59	1.60	1.92	2.11
9	0.95	1.50	1.59	1.60	1.93	2.12
8	1.05	1.62	1.74	1.75	1.94	2.12
7	1.05	1.71	1.72	1.73	2.06	2.36
6	1.35	2.60	2.77	2.78	3.57	3.91
5	4.25	8.49	8.13	8.20	10.14	10.75

TABLE 4-21 DD Code Redundancy Measure, Varying Sample Size and Number of Letters Used

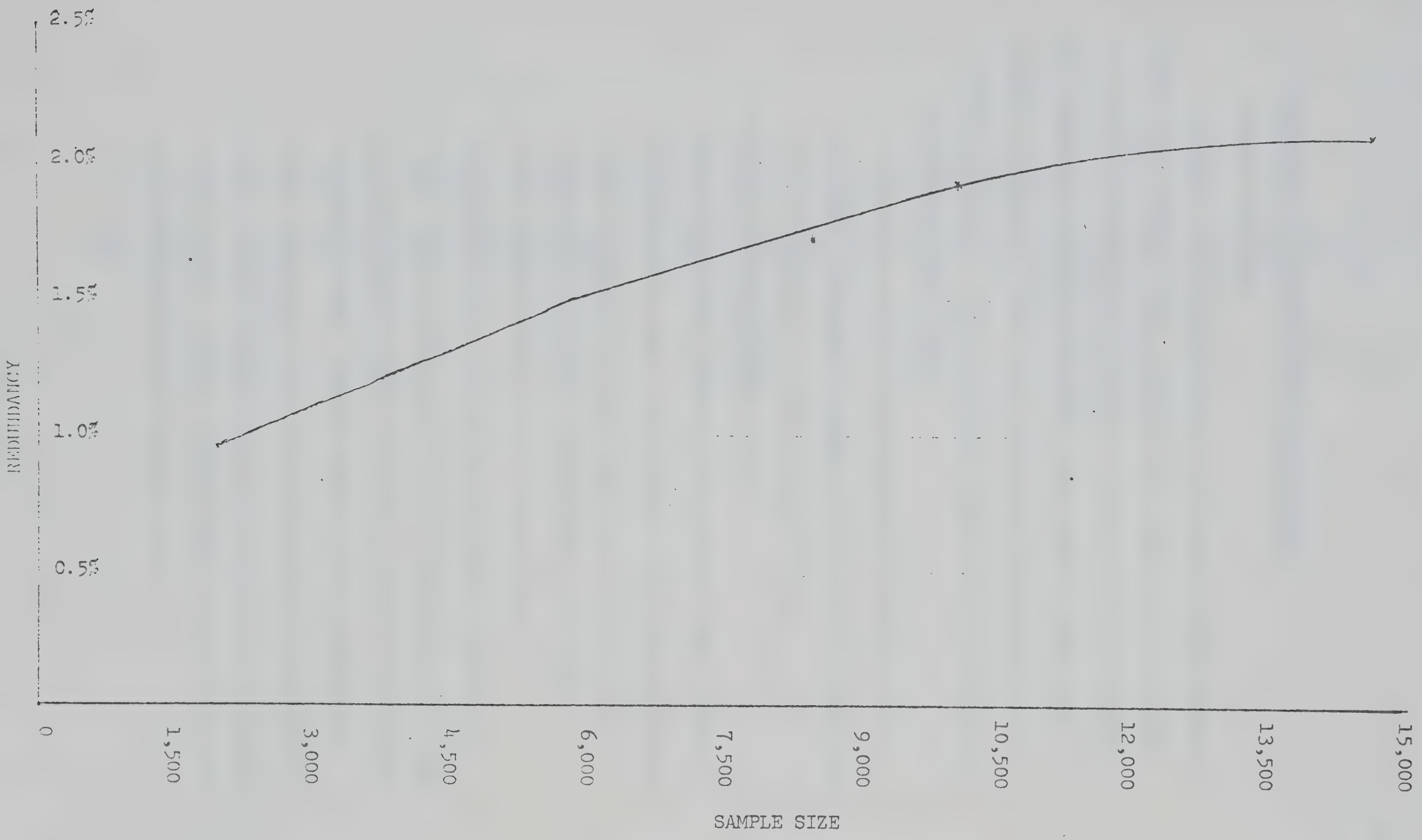


DIAGRAM 4 - 4 Redundancy vs Sample Size, Using 10 Letter Code

G. Comparison of Different Name Coding Methods

(against DD Code)

To compare the relative merits of different coding methods, three other methods were programmed with sample sizes of 10038 and 14556. The redundancy of each is measured and compared with the redundancy generated by the DD coding method (using same number of letters). Only results obtained from sample size of 14556 were tabulated. The results as compared with DD code are:

1. SOUNDEX (Table 4-2.2)

   DD Code is a superior coding method, in addition to all the other advantages discussed in Section B of this Chapter, Advantages of a System using DD Code.

2. Davidson's Method (Table 4-2 3)

   Davidson's Method appears to be superior based on just discriminating power, but because of its selection process which must start with full last name, it can not be implemented in a direct dialed fashion as DD Code.

3. Blair's Method (Tables 4-13, 4-14 and 4-2 4)

   From Table 4-22, it is clear that Blair's method has higher discriminating power than DD Code because it simply generates codes of lower redundancy. Again, by the same argument as above, Blair's Method is not suitable for direct dialed use. An interesting point is that Blair's method actually generates codes with higher discriminating power than the original alphabet when less than 7 letters are selected(see Tables 4-13 and 4-14).

TABLE 4-22 N =14556 DD Code vs Davidson's Code (4 Letters of Last Name, First Initial)

	LAST NAME	SAME CODE	SAME ALPH	DIFFERENCE SC -SA	REDUNDANCY IN PERCENT	UNIQUE CODE N -SC	UNIQUE NAME N - SA	UNIQUE CODES IN PERCENTAGE	UNIQUE NAMES IN PERCENTAGE
Davidson's		3160	1686	1474	10.126	11396	12870	78.29	88.42
DD Code		3560	1686	1874	12.874	10996	12870	75.54	88.42

TABLE 4-23 N =14556 DD Code vs SOUNDEX Code (Last Name, 4 Letters)

	LAST SAME NAME CODE	SAME ALPH	DIFFERENCE SC -SA	REDUNDANCY IN PERCENT	UNIQUE CODE N -SC	UNIQUE NAME N - SA	UNIQUE CODES IN PERCENTAGE	UNIQUE NAMES IN PERCENTAGE
SOUNDEX	12161	8351	3810	26.175	2395	6205	16.45	42.63
DD Code	11946	8351	3595	24.698	2610	6205	17.93	42.63

Letters	DD Code	Blair's
10	2.109	0.144
9	2.116	0.089
8	2.123	0.007
7	2.363	-0.268
6	3.909	-1.147
5	10.752	-1.951

TABLE 4-24 N= 14556 DD Code vs Blair's Method
Redundancy Measure (last name, varying size)

Note: negative value indicates less redundancy with respect to original alphabetic names.

In conclusion, the merits of the DD Code do not lie in its discriminating power alone, but rather in its closeness with original names. Above all, DD Code is the only coding method that can be implemented directly on a conventional telephone.

A list is produced in order to show examples of names that cause redundancy in DD code (Appendix F). For example, names like Gill and Hill are coded the same because letters G and H occupy the same key on the standard telephone dial. The use of the DD code renders the concept of 'family name' inappropriate; we may now consider a 'family class' which include all last names that receive same DD code, rather than the conventional 'family name', and all these last names may be considered 'equivalent'. The problem then is that a Mr. John Hill would be confused with a Mr. John Gill. However, this problem is no worse than having 2 of John Hill or 2 of John Gill on the file. In both cases, further information (addresses) are needed to identify them. In Sections C, E we have shown that redundancy is reduced drastically when 3 letters of the first name are used (0.048%). There is no doubt that if addresses are furnished, a person can be uniquely identified using the DD coding method.

The list in Appendix F, although long, is included for reference and for further analysis of the extent of DD code redundancy.

CHAPTER V

A PROPOSAL FOR AN AUTOMATIC TELEPHONE DIRECTORY ASSISTANCE SYSTEM

A. General System Description

The objective of this study was to get some feel of the design of an automatic telephone directory assistance system based on the concept of DD Codes. A batch model with an indexed sequential file for master file and a generic search was constructed and tested on the CDC 3170/MASTER operating system at the Northern Alberta Institute of Technology. The lessons learned from it are presented below along with a description of the model.

The system is composed of 4 major components:

a. a hardwired electronic CONVERTER that accepts dialed in enquiry through telephone sets (therefore in DD Code automatically);

b. an electronic MESSAGE-SYNTHESIZER* that generates simulated human voice output;

c. a MASTER file on direct access storage;

d. a SEARCH program that receives coded search records (in DD Code) from the CONVERTER and performs record matching on the MASTER file.

*Note: The basic idea of a Message-Synthesizer is to store prerecorded phrases on a drum type storage device. Controlled by a computer, different combinations of phrases can be synthesized into a sentence and then output through a telephone receiver. Details in Trupp's paper[1] and Winkleman's paper[3].

System Block Diagram

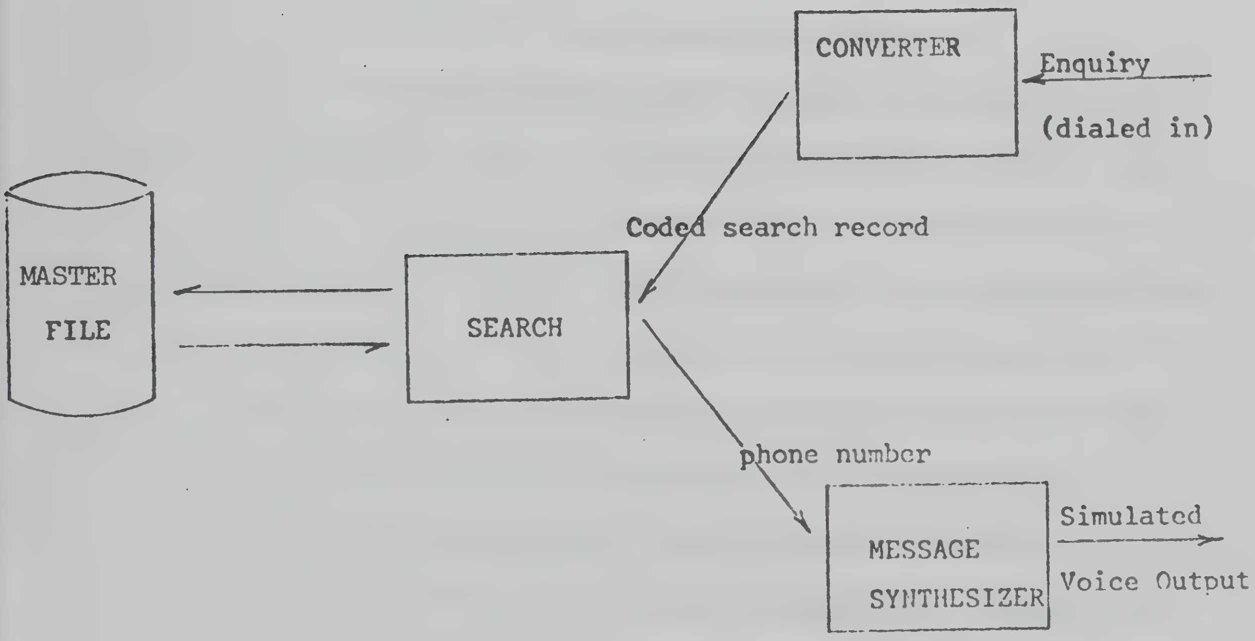


DIAGRAM 5-1

1. CONVERTER

The CONVERTER accepts dialed-in message and stores them in a buffer until all required fields are specified. There are six fields for a generalized system and five fields for a basic system. They are:

a. locality : city or district name (for a generaliazed system only);

b. class : residential or non-residential; (the latter includes government agencies and commercial listings etc.)

c. last name - first N letters of the last name. Blank filled if less than N. The value of N is to be determined for the individual implementation. (The present study tested values of N from 5 to 10, and showed that N is greater than or equal to 7)

d. first name - first M letters of the next name most frquently used (values of M from 1 to 3 were tested)

e. house number - 5 digits. e.g. 13204

f. street number - 3 characters. e.g. 103, JAS (alpha or numeric as appropriate)

Input messages are coded in numeric DD Code according to the following scheme:

1 for Q, Z, and blanks

2 for A, B, C

3 for D, E, F

4 for G, H, I

5 for J, K, L

6 for M, N, O

7 for P, R, S

8 for T, U, V

9 for W, X, Y

This, of course, is merely the standard code already in use on telephone dials, with the addition of the characters on the 1 button.

On a Touch-tone telephone set there are twelve buttons, and one of the two extra buttons (*,#) can be used as a delimiter if so desired. This is to reduce time wasted on keying blanks for filling up a field. The zero button is not needed in this application, however, it can be used as 'delimiter' for the conventional 10-key dial.

2. SEARCH - The search program accepts encoded information from the CONVERTER and a search is done on an on-line master file to find a match for the inquiry.

   a. if there is a match, it will output the telephone number through the MESSAGE-SYNTHESIZER.

   b. if there is no match, a message pertaining to this is output.

   c. if there is more than one match (i.e. given information cannot uniquely identify a record) it will output a message to ask for a more specific inquiry.

Output from either a, b, or c is passed on to the MESSAGE-SYNTHESIZER.

3. MESSAGE-SYNTHESIZER - Output from the search will cause a prerecorded human voice to be played and heard by the caller on the telephone.

4. MASTER-FILE - The system Master file contains all the telephone subscriber listings and is stored on a direct access storage device. (disk or drum or data cells) It is subdivided into two files.

   - an R-file for residential listings.
   - a B-file for all business and governmental listings.

For a more general system (used for inter-city telephone directory assistance service) a higher level of identifier can be added -- locality list, the purpose of which is to identify different regions.

B. FILE STRUCTURE, FILE ORGANIZATION AND DESCRIPTION

There are three levels of list that make up the master file for the generalized system (a basic system will have all but the Locality List).

1. Locality List: A list of cities and counties served by the system. The list is sequenced in descending order by the population size. Each record points to a record in the class list.

   Search Method: Simple sequential search

   Record Format:

KEY	STARTING ADDRESS	ENDING ADDRESS

GENERAL FILE STRUCTURE FOR AN UNIVERSAL FILE

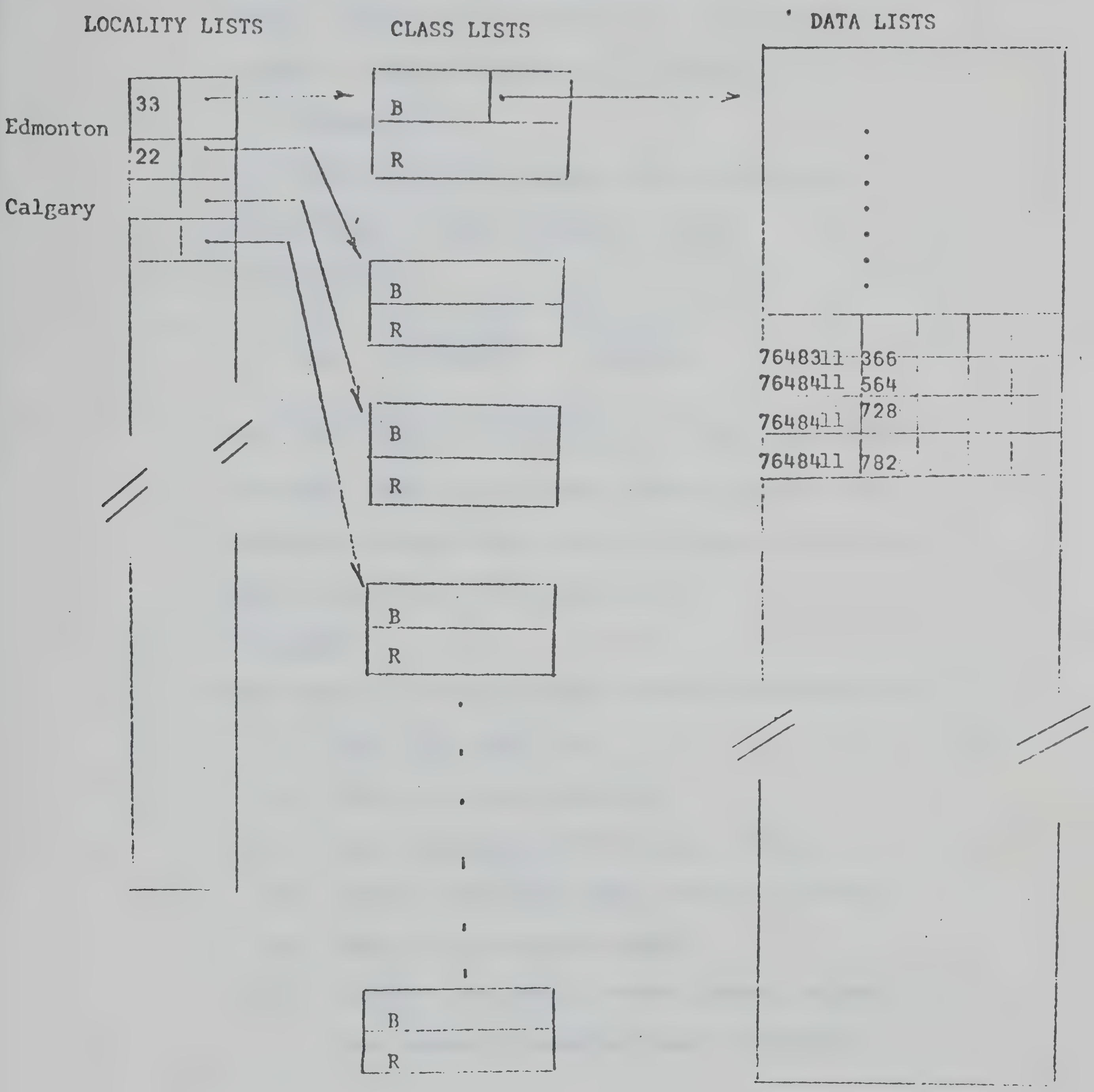


DIAGRAM 5-2

2. Class List: With each Locality, several classes are defined. From the study by Rothrock[2], a practical number of classes would be two or three*.

   a. Residential

   b. Commercial and governmental (non-residential)

   Search Method: Simple sequential search

   Record Format:

KEY	STARTING ADDRESS	ENDING ADDRESS

3. Data List: There is a data list for each class within Locality. There are two types defined, an R-file for residential listings and a B-file for business and Government listing (i.e. non-residential).

   a. <u>R-file</u>

   Each record in the R-file shall contain six fields:

   (i) last - for family name

   (ii) first - for next name given

   (iii) house - for house number (e.g., 11532)

   (iv) street - for street number (e.g., 125 Avenue)

   (v) phone - for telephone number

   (vi) pointer - for special record linkage purpose (see Chapter VI possible future development)

* A possible third class could be defined. In current practice by telephone companies, a Frequent Called Number List (FCNL) is kept for all frequently referred numbers extracted from the other two classes. Each record in the class list points to a data list.

RECORD FORMAT

KEY

Last NAME	FIRST NAME	House NUMBER	STREET NUMBER	PHONE NUMBER	POINTER

R-FILE

KEY

NAME	TYPE	HOUSE	STREET	PHONE	POINTER

B-FILE

DIAGRAM 5-3

According to Rothrock[2], of all requests on residential listings, very seldom was other information (e.g., middle initial, title) given to help increase the discriminating power. In fact, 36% of the requests can furnish only family name and next name.

b. <u>B-file</u>

In the B-file, each record also contains six fields:

(i) name - listed company name

(ii) type - business type

(iii) house - house number

(iv) street - street number

(v) phone - telephone number

(vi) pointer - future consideration

In this division, 40% of all the requests can only furnish NAME and TYPE, and 29% can supply the street name in addition to NAME and TYPE.

The emphasis of the system should be placed upon the handling of the B-file because according to Rothrock's survey of seven larger cities, over 72% of all the requests for telephone directory assistance are for business listings. This figure confirms the finding of Edmonton Telephones. Therefore, it is clear that any automatic telephone directory should stress the efficient handling of business listings.

The first four fields of both types are used as a SEARCH key; (i) is the major key; (ii) is used only if (i) alone fails to identify a unique record. Other keys are used in the same fashion.

Records are arranged in ascending order of the combined keys.

Search Method - Generic Search

a. A random search using Key (i), last name

b. sequential search using other keys

C. <u>File Creation and Maintenance</u>

1. File Creation

The creation of the master file involves three steps:

a. Conversion: Data base is originally on cards. All information is in alphanumeric code exactly like a telephone book. The card file is read and converted into the all numeric DD code. (Rule of converting as described before). The converted file is then stored on disk.

b. Sort: The disk file from (a) is read and sorted into ascending order. The choice of sorting algorithm depends on the size of file. (e.g., tournament sort). The sorted file is stored back on the disk.

c. Creation: An indexed sequential file is created from the disk file. The result is a <u>LISA</u> file again on disk.

2. File Search (Generic Key Search)

a. Random Search: A random search is made on the major key given (last name). After the first record with the major key is found, a

OPERATIONS SYSTEM FLOWCHART

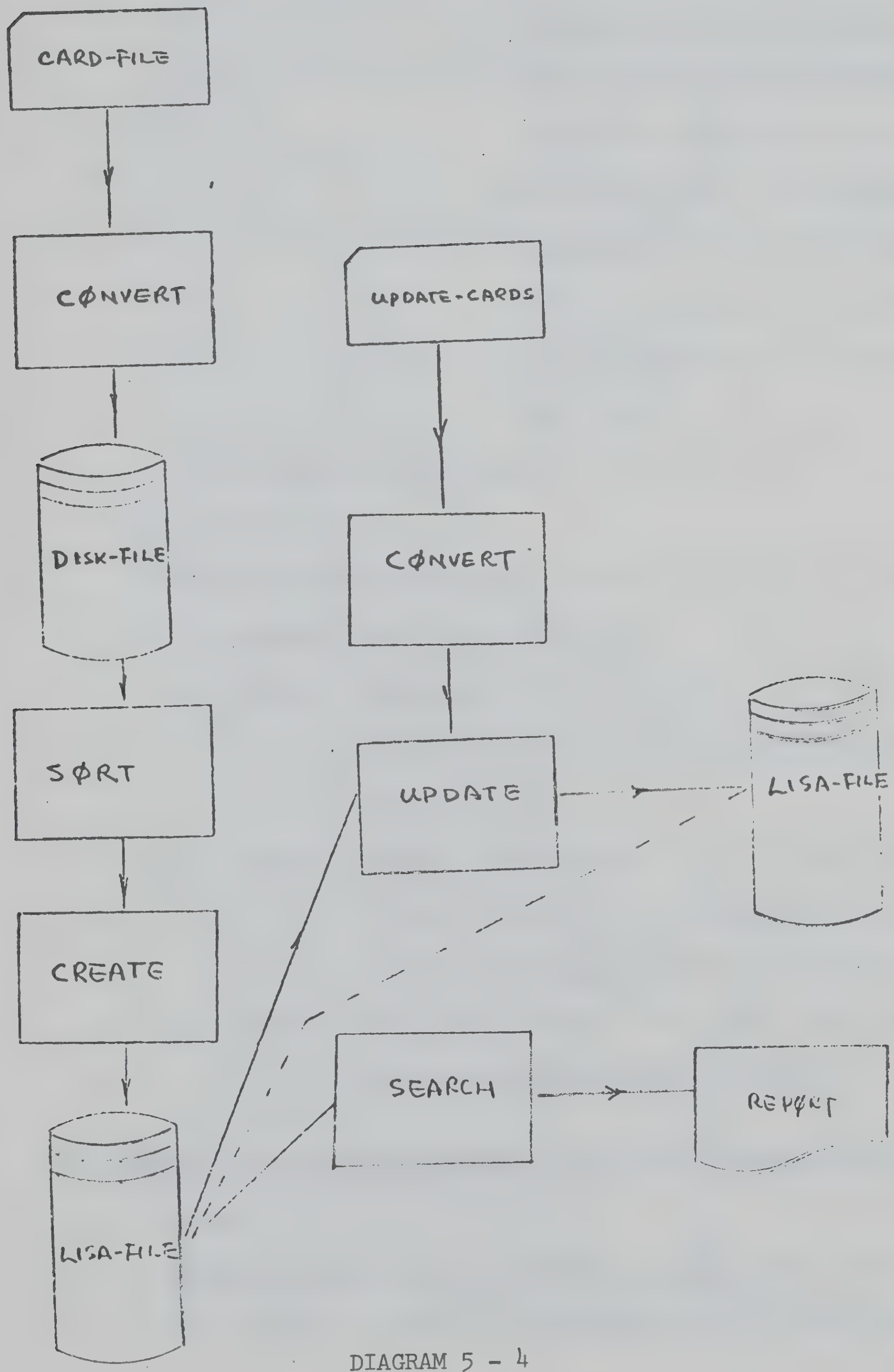


DIAGRAM 5 - 4

sequential key search is performed.

b. Sequential Search: After the successful retrieval of the first record of the given major key, requests are made by sequential read function until a major key change. If information on a certain field is not given, that particular field is masked. (No comparison will be made on that field).

c. A report on the search will be printed out.

3. File Update

There are three different functions involved in an update. All transactions from a card file.

a. Delete: Cancellation of old listings. A DELETE routine will delete record with matching key. (Major and minor).

b. Replace: Mainly changes made on old listings (e.g., change of address). A REPLACE routine will replace the <u>whole</u> record with new record.

c. Insert: Enter new listings. An INSERT routine inserts given record into the file.

D. <u>System Performance Measurement</u> (CDC Linked Index Sequential File)-LISA

1. Update

Update on master was estimated to be about 10% of the file. This becomes a crucial consideration in using this system.

With many updates, the file organization may be such that access times are greatly degraded by large numbers of overflow entries. A suggested solution would be to reorganize the file regularly (i.e., build a new edition of file).

2. Mass Storage use is the percentage of block space occupied by user's records compared to the block space available for user's records.

   There are 11 parameters to be considered:

   A: number of accesses

   B: number of buffers

   R: number of data records

   S_B: block size in words

   S_K: key size in words, fractions to be rounded to the next highest word.

   S_R: record size inwords, fractions to be rounded to the next highest word

   OB: number of overflow blocks

   IB: number of index records per block

   F: fill percentage for data block

   NB: number of data blocks

   P: total percentage of record increase during file life

Usage:

$$U = \left[1 - \frac{\left(1 + \frac{P}{100}\right) x\ R\ x\ S_R}{(NB + OB)\ (S_B - 2)} \right] x\ 100$$

$$OB = 0 \quad \text{if} \quad 100\text{-}F \leq P \text{ , otherwise } \quad OB = \frac{|100\text{-}F\text{-}P|}{100} \text{ X NB}$$

where a fractional result must be rounded to the next integer.

3. Random Record Retrieval and File Maintenance

The actual number of accesses of mass storage for retrieval or updating of a record is influenced by a number of factors:

a. the file size

b. the number of I/O buffers

c. the block size

d. the key size

e. the number of data block overflows

To determine the number of accesses, the following parameters have to be calculated:

a. $IB = \frac{S_B - 2}{S_K + 1}$ e.g., $\frac{56}{4 + 1} = 11$

(fractional value to be truncated)

b. number of data records per block (blocking factor)

$BF = \frac{S_B - 2}{S_R + 1}$ e.g., $\frac{56}{7 + 1} = 7$

c. $NB = \frac{R}{BF} \times \frac{100}{F}$ e.g., $\frac{200{,}000}{7} \times \frac{100}{375} = 38{,}000$

d. number of secondary index blocks $SIB = \frac{NB}{IB}$

e. number of primary index blocks $PIB = \frac{SIB}{IB}$

The performance of LISA in retrieving random records may be measured by the average number of mass storage accesses (A) to retrieve a record.

If $1 \leq B \leq PIB$

Then $$A = \frac{PIB - B + 2}{2} \times \frac{PIB - B + 1}{PIB} + 2 + \frac{OB}{NB + OB}$$

Or if $B \geq PIB$

$$A = 2 + \frac{OB}{NB + OB}$$

CHAPTER VI

CONCLUSION

As stated, the purpose of this report is to outline an automated telephone directory assistance system which could function without conventional human operators to provide the link between a customer and the computer.

The design is aimed at a generalized system which would serve a country or even the whole continent. However, due to a lack of resources, the experiment carried out to test the feasibility and efficiency is done only at the basic system level. Thus, there are several areas which can be considered for future development.

A. The design of a cross-country telephone directory assistance system network which serves the needs of a country; basically, all discussions above can be applied to building local systems which can then be linked together to become a "super - system". Communication networks experimented extensively at the university level in the United States. The Canadian Government is seriously considering implementing a similar system here in Canada. The possibility of and necessity for a telephone directory assistance network is definitely there. There are several points to be investigated:

1. a unified file structure for all local systems for standardization;
2. a mechanism for linking local systems (similar to a long-distance telephone call);
3. a policy on responsibility and liability for the super - system;

4. a policy on responsibility and liability for each sub-system;
5. a charging scheme to the users. First of all, it is obvious that the charging for such a service is inevitable. Secondly, charges for a local inquiry should be different from an out-of-town inquiry;
6. a profit sharing scheme among the sub-systems. When an out-of-town inquiry is made, who-should-get-what-percentage of the revenue must be determined;
7. the administration of the super-system.

As we can see, a network of this nature has great potential for other applications to serve the community. For instance, credit card agencies will find such a system very useful to help trace down debtors who move to another part of the country.

In our society, the telephone has become an integral part of our home. A telephone number could certainly be considered as a unique identification number. Perhaps, using a unique telephone number would be a step toward a unified person identification number system.

B. In the basic system, four fields (last name, first name, house number, and street number) must match ('don't know' is considered as a match) whatever is on a record before the telephone number on the record will be given. In order to maximize the chance of a match (if it is considered desirable) to warrant at least one telephone number to be given, a weighted scheme can be employed. That is, after each comparison, the degree of 'equalness' is assessed and a weight shall be assigned. After all necessary comparisons are done the record with the highest weight shall be

considered a match. Two problems arise naturally:

1. how to determine the weight to be assigned. Blair's letter weight and position weight can be consulted.
2. how many comparisons should be made. Would it be reasonable to compare all records that match last name field perfectly?

The advantage of a weighted scheme is that "minor errors" in spelling sometimes can be "overlooked" in correct retrieval of a telephone number. The obvious disadvantage is the higher probability of erroneous retrieval of a telephone number.

C. In the basic system, a sequential search on file is performed after the first record with the correct last name (accessed by a random search). It is done in this fashion because of the low frequency of occurrence of names with the same last name. A modified version would be a complete tree structure as described in the generalized system. Records with identical sub-fields (first name, house number, street number) would be linked as a list structure.

D. A simultaneous questioning-searching scheme. Search shall be carried out simultaneously with the collection of data from the caller. As each field is completed, its search will be initiated. If a unique answer is recognized before all fields of information are supplied, the system shall terminate the questioning and gives the answers to the caller. There are two disadvantages:

1. during the time a customer is keying the information, it ties up the terminal;
2. erroneous specification by the caller results in wasteful search time,whereas the system described in this report would only start searching when a "send" signal is given by

the caller after he correctly keyed-in all information.

E. Creation of an inverted index on telephone numbers for the master file. This would help in the aspects of accounting and charging. Also, it provides record linkage capability for other applications such as law enforcement.

F. Creation of a list of newly assigned telephone numbers corresponding to the old telephone number of the same customer. This would enable a caller to find out a new telephone number with insufficient knowledge of the regular search information (first name, etc.). In the future, he can dial the correct number.

In conclusion, all the above suggested future developments are geared to a more versatile and powerful telephone directory assistance system. Whether they are applicable and feasible for implementation depends largely on the needs of each individual local telephone company. We may consider the following:

1. what level of service the local telephone company wants to achieve;
2. how much the local telephone company is willing to spend in order to attain that level of service;
3. is the time constraint an important factor (e.g. Edmonton Telephones find they must have a new system to replace the old manual system by 1974);
4. the area of standardization of file and data structure must be examined carefully to ensure future compatibility with other systems to make a network.

As it is generally conceded that no system is perfect, an open-

ended system would definitely be more dynamic and adaptable to new requirements and changing demands. In this era of rapid technological advancements, we cannot afford to let the field of information processing stay behind. Information is wanted and wanted fast. It is certainly unwise to try to design a perfect system at the expense of time. Instead, a simple functional system that is easy to upgrade and flexible enough to utilize future computing hardware and software facilities would prove itself to be most desirable and economical.

Looking ahead into the future, with the advent of this direct dialed system, potentially every household with a telephone set has a computer linked terminal. Prophecies of housewives shopping without leaving home, or of a man doing business with a computer via his telephone set, can be fulfilled in the near future. What is more, conversion to this 'terminal age' is easy and painless as the telephone is readily available in most households and offices nowadays. Direct Dialing is just an step toward this goal, but the implications are very far-reaching and significant.

REFERENCES

1. R. D. Trupp, Computer-controlled Message Synthesis, Bell Lab Record, June/July, 1970.

2. H. I. Rothrock, Computer-assisted Directory Search, Doctoral thesis, University Microfilms, 1970.

3. W. A. Winckelmann, Automatic Intercept Service, Bell Lab, May, 1968.

4. E. Fredkin, "Trie Memory", Comm. ACM, 3L490, 1960.

5. Business Equipment Manufacturers Association, ANSI Standard, Identification of Individuals for Information Interchange, June, 1970.

6. Charles R. Blair, "A Program for Correcting Spelling Errors", Information and Control, vol. 3, 60-67, 1960.

7. G. A. Miller and E. A. Firedman, "The Reconstruction of Mutilated English Texts", Information and Control, vol. 1, 38 - 55, 1957.

8. C. P. Bourne and D. G. Ford, "A Study of the Statistics of Letters in English Words", Information and Control, vol. 4, 48 - 67, 1961.

9. L. N. Korloev, "Coding and Code Compression", J. ACM, vol. 5, 328-330, 1958.

10. Remington-Rand, SOUNDEX - a foolproof filing system based on sound rather than spelling, Brochure LBV809R1, undated.

11. Remington-Rand, Searching aids for alphabetic and SOUNDEX files, Brochure LBV 440A Rev. 1, undated.

12. B. W. Taunton, "Name-code - a Method of Filing Accounts Alphabetically on a Computer", Data Processing, vol. 2, 23 - 24, 1960.

13. L. Davidson, "Retrieval of Misspelled Names in an Airlines Passenger Record System",Comm. ACM, vol. 5, 169 - 171, 1962.

14. L. Hogben, M. M. Johnstone, and K. W. Cross, "Identification of Medical Documents", British Medical Journal, (Apr. 1948), 632-635.

15. IBM, Coding Methods, IBM data processing techniques series, GF20 - 8093 - 0.

16. IBM, Identification Techniques, IBM data processing techniques series, GC20 - 1707 - 0.

17. C. Auger, A. Meidels, P. Ardouin and G. Kirouac. "Patient Identification for Computer Data Linkage", Canad. Med. Ass. J., Dec. 13, 1969, vol. 101, 747 - 749.

18. E. K. Huffman, Manual for Medical Record Librarians, 5th ed., p. 251-255, Physician Record Company, Berwyn, Ill., 1963.

19. E. D. Acheson and J. G. Evans, "The Oxford Record Linkage Study: a Review of the Methods with Some Preliminary Results", Proc. Roy. Soc. Med., vol. 57, 269 - 274, 1964.

20. Committee on the implications of Record Linkage for Health-Related Research, Health Research Uses of Record Linkage in Canada, a report to the Medical Research Council of Canada, October 1968.

21. H. B. Newcombe and J. M. Kennedy, "Record Linkage - Making Maximum Use of the Discriminating Power of Identifying Information", Comm. ACM, 5, 563 - 566, 1962.

22. G. A. Miller and E. B. Newman and E. A. Freidman, "Length Frequency Statistics for Written English", Information and Control, vol. 1, 370 - 389, 1958.

23. H. B. Newcombe, A. P. James, and S. J. Axford, Family Linkage of Vital and Health Records, Atomic Energy Canada Report: AECL - 470, CRB717, July, 1957.

24. H. B. Newcome, J. M. Kennedy, S. J. Axford, and A. P. James, Automatic Linkage of Vital Reocrds, Science 130, 954 - 959, 1959.

READING LIST ON COMPUTER MESSAGE SYNTHESIS

1. L. Rabiner, Speech Synthesis by Rule: An Acoustic Domain Approach, Bell System Tech., 47, 17-37 (1968)

2. B. Gold and L. Rabiner, "Analysis of Digital and Analog Formant Synthesizers", IEEE Trans. on Audio and Electroacoust,Vol. AU-16, No. 1, March 1968.

3. L. Rabiner, "Digital Formant Synthesizer for Speech Synthesis Studies", J. Acoust. Soc. Am. vol. 43, pp. 822-828, 1968.

4. L. Rabiner, R. W. Schafer, J. L. Flanagan, Computer Synthesis of Speech by Concatenation of Formant-Coded Words, Bell System Tech., 50, 1541-1558 (1971)

5. L. Rabiner, "A Model for Synthesizing Speech by Rule", IEEE Trans on Audio and Electroacoust, vol.AU-17, no. 1, March 1969.

6. J. L. Flanagan, Speech Analysis, Synthesis and Perception, Academic Press Inc., New York, 1965.

7. J. L. Flanagan, "Recent Studies in Speech Research at Bell Telephone Labs (II)", Proceedings of International Congress of Acoustics, 5th, Liege (1965), Paper A22.

8. R. W. Schafer and L. R. Rabiner, "System for Automatic Analysis of Voiced Speech, J. Acoustics Soc. Amer., 47 (Feb. 1970), pp. 634-648.

9. J. L. Flanagan, C. H. Coker, L. R. Rabiner, R. W. Schafer and N. Umeda, "Synthetic Voices for Computers", IEEE Spectrum, 7, (Oct. 1970), pp. 22-45.

10. L. R. Rabiner, L. B. Jackson, R. W. Schafer and C. H. Coker, Digital Hardware for Speech Synthesis, Seventh International Congress on Acoustics, Budapest, August 1971.

11. W. J. Strong, "Machine-aided Formant Determination for Speech Synthesis", J. Acoustic Soc. Amer. Vol. 41, pp. 1434-1442, 1967.

12. L. L. Beranek, Acoustic Measurements, New York, Wiley, 1965.

13. L. R. Rabiner, H. Levitt and A. E. Rosenberg, "Investigation of Stress Patterns for Speech Synthesis by Rule", J. Acoust. Soc. Am., vol. 45, pp. 92-101, 1969.

APPENDIX

USER INSTRUCTIONS

1. For a long distance inquiry, a user should dial 1 first and then dial the area code. (e.g., 604 for B. C.).
2. Dial for automatic directory assistance (Assume still 411).
3. Wait for a ready tone.
4. Enter a sequence of four items of information (Called fields) Each field separated by pushing a delimiter button.
    a. last name (or finding name for a business listing); use only the first 7 letters.
    b. first name (or business type if it is a business listing); use only the first three letters.
    c. house number; maximum: five letters or digits.
    d. street number (or street name); maximum: 3 letters or digits

    If any item of information is less than the allowed maximum length, the user just dials blanks to fill up the field, or uses the field terminator.
5. When all four items are entered, the user shall hear the answer from the telephone.
6. The answer may be in one of three forms:
    a. one telephone number if there is one and only one "find".
    b. an apology if there is no "find".
    c. if there is more than one "find", several numbers will be given and the user may try them. (The maximum number of telephone nos. to give on one request can be determined by each individual installation).
7. The line is disconnected.

Note:

a. The efficient use of the system depends on the accuracy and completeness of the search information provided by the users. Therefore, it is the user's responsibility to provide accurate and complete information to ensure successful information retrieval (e.g. correct spelling of names).

b. Often a user may be uncertain about a field. He should make separate inquiries with each change on the search information. Identical search information would only result in an identical answer.

c. There shall be an emergency number for "desperate cases". That is, a human operator shall come to the user's assistance if he dials this "emergency" number.

DIALING RULES

1. All the letters (except Q and Z) have an equivlent number code as shown on a standard telephone dial disk (also the same as a touch-tone telephone panel). Therefore to dial a letter, just dial the equivalent number on the dial. (e.g., the number for D, E, or F, is 3)
2. Dial 1 for Q, Z, or blanks.
3. Numerals are unchanged. (e.g. 3 is still 3)

EXAMPLES:

a. (i)	last name (7)* BLACK	2522511
(ii)	next name (3)* JOHN	564
(iii)	house number (5)* 12345	12345
(iv)	street number (3) 104 St.	104
b. (i)	last (7) ANDERSON	2633776
(ii)	next name (3) WILLIAM	945
(iii)	house number (5) 8204	82041
(iv)	street number (3) Jasper Ave.	527

* Note: A delimiter can be implemented to separate fields, (special key on the new touch-tone keyboard). User may dial as many characters as he wishes and closes each field with the delimiter (e.g. *). Fields that are too long shall be truncated while fields that are short shall be blank filled.

RULES OF SOUNDEX

1. The first letter of a surname is retained in its uncoded form and is termed a prefix.

2. Other letters of the surname are assigned code numbers as follows:

B, P, F, V	= 1
C, G, J, K, Q, S, X, Z	= 2
D, T	= 3
L	= 4
M, N	= 5
R	= 6

A, E, I, O, U and Y are not assigned a code but serve as 'separators' (see below); W and H are ignored entirely.

3. The second of a pair of consecutive identical digits may be retained as part of the code only if the corresponding consonants are separated by a vowel or Y. The rule applies in a similar fashion to a digit which would follow a prefix letter having the same code.

4. The coding stops when three digits have been obtained. If the coding yields less than three digits, zeros are used to complete the code.

FACTORS INFLUENCING CHOICE OF IDENTIFYING INFORMATION

MRC (Medical Research Council) Report #3 (1968)

Appendix II: Record Linkage in Canada

1. Social Insurance Number (SIN)
2. Full Name
3. Mother's maiden name
4. Day, month and year of birth
5. Place of birth
   (city, province if in Canada, or city and country if not in Canada)
6. Sex

Possible addition: woman's maiden name, if husband's name were used.

Error: a. failure to link records that ought to be linked.

b. linkage of records that ought not to be linked.

A measure of 'redundancy' - make multiple comparison of records.

Acceptable error level depends on particular project.

Feasible linkage requires:

1. name: first, middle initial, last.
2. mother's maiden name.
3. date of birth
4. place of birth
5. sex

6. marital status

Additional information;

1. first name of spouse
2. maiden name if married woman
3. place of residence (city, province)
4. father's first name
5. mother's first name.

Date of registration of the record, used:

1. entry on record
2. parameter in the record.

Best solution is a universal identification number.

SWEDISH GOVERNMENT UID SCHEME E - 1

Personal Identification Number

The personal identification number has ten digits with a dash after the first six. It consists of the following three parts:

a) birth date (6 digits)
b) birth number (3 digits)
c) control figure

|3 8 0 4 2 5| - |6 6 5|3|

a) The birth date is indicated by six digits in the following order:

two last digits of year of birth
month
day

|3 8|0 4|2 5|

b) The birth number has three digits, odd for men and even for women. It can be any of the numbers 001 - 999. Persons born on the same day shall have different numbers.

c) The control figure is added to the birth number and can as a rule be used to test that no wrong figures have been given in birth date and birth number. It is calculated in the following way:

1. All the single digits in birth year, month and day, and birth number are multiplied alternately by 2 and 1

```
3 8  0 4  2 5  6 6 5
2 1  2 1  2 1  2 1 2
```

6,8, 0,4, 4,5, 12,6,10,

2. Add the received figures. Note that 12=1+2

6+8+0+4+4+5+1+2+6+1+0=37

3. The last digit of the sum is decuted from the number 10.

10-7=3

4. The figure arrived at is the control figure. If this figure is 10,

the control figure will be 0.

Birth numbers were introduced in Sweden on January 1, 1947. The control figure was added on January 1, 1968.

Every person registered for census in Sweden shall have a personal identification number, regardless of citizenship. Such numbers with the exception of the control figure, have been given to all persons registered in Sweden on January 1, 1947, and to all who thereafter have been registered.

In some instances personal identification numbers are given to persons not census registered in Sweden, for example, persons doing military service in Sweden or who pay taxes in Sweden are registered with the Swedish Social Insurance Service.

F-01

Partial

LIST OF DIFFERENT NAMES*

THAT GENERATES SAME DD CODES

COMMON NAMES** WERE INDICATED

*Based on 14556 students records

**Common names table[16]

1. Base on analysis of U.S Social Security[16] records 117,358,888

2. Includes 1586 names representing 48% of persons on record

3. Rank number indicates ranking of 121 most common names

```
        LIST OF DIFFERENT NAMES NAMES HAVING SAME CODE    F-02

    NAME CODE  LAST NAME
    1238629754 ZADUNAYSKI
    1238629754 ZADVNAYSKI

    2222511111 CABAJ
    2222511111 CACAK

    2246111111 BAIN
    2246111111 CAIN          ***COMMON 2246111111 CAIN

    2253711111 ACKER
    2253711111 BAJER
    2253711111 BAKER         ***COMMON 2253711111 BAKER     RANK= 31
    2253711111 BALES

    2266261111 BANMAN
    2266261111 CANNAN

    2272611111 ABRAM
    2272611111 BARAN

    2273111111 BARD
    2273111111 CARD
    2273111111 CASE          ***COMMON 2273111111 CASE

    2273911111 CARDY
    2273911111 CAREY         ***COMMON 2273911111 CAREY

    2276611111 BARON
    2276611111 CARON

    2277111111 BARR          ***COMMON 2277111111 BARR
    2277111111 BASS          ***COMMON 2277111111 BASS
    2277111111 CARR          ***COMMON 2277111111 CARR

    2277388111 BARRETT       ***COMMON 2277388111 BARRETT
    2277388111 BASSETT       ***COMMON 2277388111 BASSETT

    2277661111 BARRON        ***COMMON 2277661111 BARRON
    2277661111 CARSON        ***COMMON 2277661111 CARSON

    2278371111 CARTER        ***COMMON 2278371111 CARTER    RANK= 41
    2278371111 CASTER

    2283711111 BATES         ***COMMON 2283711111 BATES
    2283711111 BAUER         ***COMMON 2283711111 BAUER

    2286611111 ACTON
    2286611111 CATON

    2325111111 BECK          ***COMMON 2325111111 BECK
    2325111111 CEAL
```

```
      LIST OF DIFFERENT NAMES NAMES HAVING SAME CODE    F-03

NAME CODE  LAST NAME

2355371111 BEKKER
2355371111 BELLER

2376371111 BERNER
2376371111 BERNES

2453111111 AGLE
2453111111 AHLF

2473111111 AIRD
2473111111 BIRD        ***COMMON 2473111111 BIPD

2474111111 BIRI
2474111111 BISH

2476111111 BIRN
2476111111 BIRO

2477355111 BIRRELL
2477355111 BISSELL
2477355111 CISSELL

2524711111 BLAIR       ***COMMON 2524711111 BLAIR
2524711111 BLAIS

2527511111 BLASK
2527511111 CLARK       ***COMMON 2527511111 CLARK      RANK= 18

2647837811 BOISUERT
2647837811 BOISVERT

2664611111 BONIN
2664611111 COMIN

2665391111 CONLEY      ***COMMON 2665391111 CONLEY
2665391111 COOLEY      ***COMMON 2665391111 COOLEY

2666111111 BONN
2666111111 BOON
2666111111 COMM
2666111111 COON

2666371111 BOOMER
2666371111 COONES

2667111111 AMOS        ***COMMON 2667111111 AMOS
2667111111 BOOS

2672411111 BOSCH
2672411111 CORAH
```

```
      LIST OF DIFFERENT NAMES NAMES HAVING SAME CODE    F-04

NAME CODE  LAST NAME

2675263111 BOPLAND
2675263111 COPLAND

2677111111 BOSS
2677111111 COPP

2682511111 ANTAL
2682511111 BOUCK

2691111111 COX          ***COMMON 2691111111 COX        RANK= 64
2691111111 COY

2692611111 BOYCO
2692611111 COWAN        ***COMMON 2692611111 COWAN

2693111111 BOWD
2693111111 BOYD         ***COMMON 2693111111 BOYD
2693111111 COWE

2693611111 BOWEN        ***COMMON 2693611111 BOWEN
2693611111 COWEN

2693711111 BOWER        ***COMMON 2693711111 BOWER
2693711111 BOWES
2693711111 BOYER        ***COMMON 2693711111 BOYER
2693711111 BOYES
2693711111 COYES

2695371111 BOWLES       ***COMMON 2695371111 BOWLES
2695371111 COWLES

2695391111 BOWLEY
2695391111 COWLEY

2724411111 BRAGG        ***COMMON 2724411111 BRAGG
2724411111 CRAGG
2724411111 CRAIG        ***COMMON 2724411111 CRAIG

2726311111 BRAND
2726311111 CRANE        ***COMMON 2726311111 CRANE

2726611111 ARBON
2726611111 BRANN

2763811111 ARNDT
2763811111 CROFT

2766537111 BROOKER
2766537111 CROOKES
```

B30069